UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO FIN DE MÁSTER

EXPERIMENTAL STUDY OF SOCIO CHATBOT USABILITY

Máster Universitario en Investigación e Innovación en Inteligencia Computacional y Sistemas Interactivos

Autora: Ranci Ren

Directores: Silvia Teresita Acuña Castillo
John Wilmar Castro Llanos

Madrid
Julio de 2019

Acknowledgement

I would like to express my gratitude to my tutor Dr. Silvia Teresita Acuña for her valuable and constructive suggestions and significant supports during this research work. I did learn a lot from her, she spent time instructing me how to write a paper, how to collect data and how to organize the research process. I thank her wholeheartedly, not only for her tremendous academic support, but also for giving me so many wonderful opportunities. I attribute the level of my Master's degree to her encouragement and effort and without her this dissertation, too, would not have been completed or written. I could not have imagined having a better tutor and mentor during my master study.

Besides, I would like to express my hugely appreciative to Dr. John Wilmar Castro for sharing his knowledge on this area, for his encouragement, insightful comments and for his guidance and his great support in the development of this work and my whole master's program.

Furthermore, I would like to thank Dr. Edison Gonzalo Espinosa Gallardo for his help, time and collaboration from Ecuador in such a generous and disinterested way. I also want to thank Andrea for valuable help during the experiment. Her willingness to give her time and share her knowledge so generously has been very much appreciated. Also, I like to thank the participants in the survey, who have willingly shared their precious time during the process of experimenting. Finally, but by no means least, I would like to thank my mother for almost unbelievable support.

Abstract

Background: The use of chatbots especially with ability of natural language processing (NLP) has increased considerably in recent years. These chatbots are being used in different areas and by a wide variety of users. Due to this variety of users, it is essential to incorporate usability in the development of chatbots.

Objective: The objective of this research is to firstly identify the state of the art of the chatbots usability and the applied human-computer interaction (HCI) techniques and analyze how to evaluate the chatbots usability, secondly to evaluate the usability of chatbot SOCIO (which helps user to create class diagram), as well as the quality of the class diagram realized.

Method: For the aim of this research, the literature was first reviewed for the purpose of specifying the state of the art in the usability of chatbot. To this end, a systematic mapping study (SMS) was be conducted through Scopus, ACM Digital Library, IEEE Xplorer, SpringerLink and ScienceDirect using a predefined search strategy. In addition, the articles and the results were analyzed. Finally, from the point of view of 18 teams of three members each, the usability of the web application CREATELY and chatbot SOCIO as well as the quality of the class diagrams obtained at using these tools were compared by executing a within-subjects cross-over design experiment. Each of 18 teams realized two class diagrams about a college and an online store by using chatbot SOCIO and CREATELY.

Results: The search retrieved 170 papers and 19 are retained as primary studies. There are few papers reviewed the chatbots usability. A proposal of usability of chatbot is proposed. Compared with CREATELY, usability of the chatbot SOCIO is evaluated in aspects of efficiency, effectiveness, satisfaction and quality. The results of this experiment indicate that the chatbot SOCIO has a positive effect on the effectiveness, efficiency and satisfaction of the participants when they create the class diagrams, as well as its quality.

Conclusions: We categorized according to four criteria: usability techniques, usability characteristics, research methods and type of chatbots. The chatbots usability is an emerging field of research, where the published studies were mainly survey, usability test, and experiments. The results of the experiment executed indicate that chatbot SOCIO performed better than CREATELY in aspects of effectiveness, efficiency and satisfaction when team making the class diagram as well as the quality of class diagrams.

Keywords: Chatbot, Chatbot Usability, Systematic Mapping Study, Usability Techniques, Usability Characteristics, Class Diagram

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

This study is framed in the areas of chatbots and usability, which raises a Systematic Mapping Study (SMS) on chatbot usability and an experiment for evaluating the usability of a specific chatbot called SOCIO. In this chapter, firstly, the research topic is described in general. Secondly, it deals with a research area. Thirdly, the research problem, the solution and the process of a control experiment executed are introduced respectively. Finally, the work structure and contribution of our work is presented.

1.1. Overview

With the worldwide popularization of computer science, network information resources have become the dominant channel for people to obtain information. As an effective internet platform communication channel, chatbots play an important role in information resources navigation and information retrieval in the current network environment and has penetrated into our lives [36].

Chatbots are defined as computer programs with a textual or voice interface, based on natural language [19]. Presently, chatbots have obtained popularity owing to their high engagement affair with users and permanent availability as Software-as-a-Service (SaaS) [38]. They are specifically designed to make user interaction as natural as possible, and they have received extensive attention from academia and industry in recent years. They represent a faster and more natural way to access information, in the near future they will also be an important crucial factor in realizing artificial intelligence as well.

Usability is defined as the degree to which a program can be used to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a specified context of use [17]. Usability is a critical aspect in interactive software systems and it is essential to incorporate usability in chatbots as well as to improve user experience [45].

Chatbots are become pervasive and are used in many areas, such as bookings of all sorts of services, to obtain medical advice and for online shopping [19][30][36]. The multiple uses and benefits of chatbots explain their strong growth in terms of users, satisfaction and saving resources. It is expected that the number of users will grow in the US by 23.1% [5]. Although the market is still beginning to take shape (compared to the number of websites, the number of bots is still not large) it is estimated that the market size will expand massively [36].

Many universities and commercial companies have put into use chatbots interacting with mature systems. At the commercial level, Facebook messenger already has more than 300,000 chatbots in use [5]. This makes downloading and installing new apps

unnecessary, and the use of smartphones allows for personalization possibilities [27]. Further, the use of chatbots can be more cost-effective than human-assisted support [24]. Some universities, such as the University of Murcia (Spain) launched systems to help students in the pre-registration and enrollment process, which has served 4,622 users, has received 38,795 messages and 13,227 conversations [25]. Even some companies are building chatbots independently (e.g., Microsoft is promoting the "conversation as a platform") to support a variety of media, from Skype to search [43].

The objective of this work is to identify the state of the art of the chatbots usability and the applied human-computer interaction (HCI) techniques by a systematic mapping study and to analyze how to evaluate the chatbots usability by executing an experiment. In this work, we report our methodology and results and provide recommendations on how developer should adopt the chatbot SOCIO to aid potential future adopters.

1.2. Research Area

1.2.1. A Brief History of Chatbots Development

Chatbots are not an emerging concept. Research on dialogue systems can be traced back to the 1950s, when Alan M. Turing posed the question "Can machines think?" and proposed the Turing test as a criterion for judging whether the machine has intelligence [22]. Weizenbaum's development of ELIZA at MIT in 1960, can be considered the first dialogue system [19]. The ELIZA chatbot took the form of a psychotherapist who answers questions for users. It just took keywords from a user's input and presented a related question as the response [34].

Afterwards, the next chatbots heavily followed ELIZA's approach with slight additions till Wallace created A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) in 1995 [38]. Unlike ELIZA, the ALICE chatbot used a pure pattern matching method and embedding the artificial intelligence markup language (AIML), which allowed for more sophisticated conversation [34].

Lately, the advances in natural language processing (NLP) have boosted the raise of many frameworks to build chatbots. For instance, chatbot SmarterChild which is launched in 2001, it not only made conversations with users, but also performed types of useful functions, instant accessing to news, sports scores and much more [3].

In recent years, chatbots have been widely used in our daily life. Many chatbots are built natively into operating systems. In 2010, Siri, the first mainstream assistant released. Users are able to interact with chatbots in natural language via text or voice. After that, Microsoft's xiaoice, Amazon's Alexa, Google Home, etc. appeared one after another. Additionally, nowadays there are hundreds of chatbot platforms which help users develop their own chatbot easily, like Dialogflow (http://dialogflow.com), Flow XO (https://flowxo.com/) and IBM Watson (https://www.ibm.com/watson) [39].

In April 2016, Facebook announced that other companies could collaborate with them using the Facebook messenger app and they could integrate their bots into the existing app. That means that any company can ride the wave of Facebook Messenger's success and huge audience to use a chatbot to engage with the customer in any way they want. Since then, tens of thousands of chatbots have been created that tie into the service.

In 2019, a number of major enterprises including British Airways, LinkedIn and Starbucks have indicated their support for chatbot development and expressed interest in applying them throughout their operations [14].

As this short summary of the chatbot field shows, various improvements have been made since the early days of NLP. Despite this, chatbots' learning capabilities are presently not always sufficient to fulfill natural-feeling conversations and they still impose a learning curve on users.

1.2.2. Usability on Chatbots Development

Usability has been a fundamental concept for HCI area. In recent years, the use of chatbot has been growing these decades, chatbots has become an important component that are involved in almost every industry [1]. Considering the statistic from Gartner, artificial intelligence will amount for 85% of customer relationships by 2020.

In the 1980s, with the reduction in computer purchase costs, more and more computer users just had basic training on applications software and operating systems or none at all. It is apparent that computers were too hard to use for almost all users, even sometimes unusable at all. The concept of "Usability" emerged then, with the purpose of developing the interactive software that would be used by everyone [46]. In the 1990s, Jakob Nielsen proposed 10 general usability heuristics for evaluating interaction design [32]. Crafting a compelling, delightful chatbot experience is going to be a key differentiator between the chatbots that see adoption and those that don't. Nielsen's heuristics provides a great benchmark to point us in the right direction [7]. The growing importance of chatbot has led researchers to study how chatbot processes differ from traditional processes. Følstad and Brandtzaeg proposed the main transition for chatbot can be summarized as follows [13]:

- Conversations as the object of design.
- The need to move from user interface design to service design.
- The need to design for interaction in networks of human and intelligent machine actors.

No matter what form of learning the chatbot uses, chatbots' main purpose is to streamline interaction between people and services. The way in which they are utilized can vary between facilitating human communication to completely replacing the need for humans to communicate with each other at all [41]. Chatbots communicate and perform basic tasks such as answering questions or placing product orders, they can be used for services or as a marketing tool for engagement. Chatbots combine the ability for scale and personalization, they can also provide content, facilitate a process, or connect with users. Platforms like Facebook, kik and WhatsApp have a combined user-base of more than one billion. Chatbots can be leverage on these apps. Chatbots will replace the search window. Chatbots streamline simple acts for customers by making it easier and more convenient to communicate with brands.

In the commercial world, one of the strongest appeals for chatbot usage is ability to reach a lot of people. Prime candidates for chatbot applications are finance, travel and retail. Chatbots become indispensable partners. Their growth and popularity are motivated by at least three different factors. First, there is the hope to reduce customer-service costs by replacing human agents with bots. Second, the success of conversational-based systems like WeChat has put forward the idea of chatbots as an interaction channel with businesses

and services, intended to supplement existing channels such as the mobile web and mobile apps. Last, the popularity of voice-based intelligent assistants such as Alexa and Google Home has pushed many businesses to apply them at a smaller scale [14].

But so far chatbots still have the problem of replacing human contact. Although they have the ability to recognize and use millions of words to give specific answers, they fall short when it comes to establishing the sort of deep and natural contact that occurs between human beings [6], especially when it comes to complex situations requiring responses that demonstrate human features, such as empathy.

While chatbots are becoming increasingly popular, that doesn't mean one chatbot can work for everyone. 60% of chatbot users are between 13 and 19 years old, with more females than males. The service seems particularly suited for Millenials and Gen Zers, who grew up using many of these on-demand technologies, but it can be difficult for brands to reach other, older demographics through chatbots. When developing a bot, companies need to be aware of their target customers and who they are trying to reach with a new technology.

1.3. Research Problem

On the one hand, in the area of the HCI, "Usability" refers to the ease of access and/or use of a product or website. It's a sub-discipline of user experience design. Although user experience design (UX Design) and usability were once used interchangeably, we must now understand that usability provides an important contribution to UX; however, it's not the whole of the experience. We can accurately measure usability [45].

On the other hand, every product or website should be easy and pleasurable to use, but designing an effective, efficient and enjoyable product is hardly the result of good intentions alone. Only through careful execution of certain usability principles can the developer achieve this and avoid user dissatisfaction, too. As chatbots have been used in our daily life, how to evaluate the usability in chatbots needs to be investigated which helps to improve the user experience.

In the work of Novick and Rodríguez [33], they did a usability evaluation of a video game bot, they found out that studies of evaluation of the usability have, with very limited exceptions, not extended to empirical evaluation through user studies. Rather, the research literature of usability has tended to focus on heuristic evaluation, and the research literature on user-centered evaluation has tended to remain at the theoretical level.

There are two research problem considered in the present research work. The first problem is how to incorporate usability in chatbots and the use of usability techniques in chatbots development. The second problem is how to obtain empirical evidence about the usability of the chatbot SOCIO that permits the creation of the class diagrams in a collaborative way.

1.4. Solution Approach

Usability is a critical aspect in interactive software systems and it is essential to incorporate usability in chatbots, to improve user experience. To do this, we studied the usability techniques, characteristics and research methods of the chatbot. This incorporation should be understood as the adaptations required by the techniques to be incorporated into chatbot developments. In addition, an experiment was executed that

could be incorporated into real chatbots projects and determine which adaptations should be considered for that purpose. Finally, its application in chatbot should be validated.

The solutions of the present problem are: First, the criteria selected for usability evaluation in the chatbots development process are: usability techniques, usability characteristics, research methods and type of chatbot. There are some works that have done to evaluate usability of some specific chatbots (see Appendix B) during the product development process. Based on those real cases, we conducted a systematic mapping study to realize how existing works evaluated to usability of chatbots, combined our acknowledges in the HCI area to analyze evaluation works of chatbot usability which were incorporated into chatbots development processes. Second, determine the feasibility of incorporating these usability evaluations into real chatbot development.

To carry out this work, previously, we have made the revision of the publications related to usability of chatbots. For this, we use a review process known as the Systematic Mapping Study (SMS). An SMS allows a review of the literature on a particular area of interest [21]. The SMS aims to answer the research question: What is the current state of usability of chatbots?

The search was carried out in five databases (DBs): IEEE Xplore, ScienceDirect, ACM Digital Library, SpringerLink and Scopus. The search was selected as the start date of January 2014 and the end date of October 2018. It is clear that, the concept of Chatbots has been proposed for many years, the number of publications started to grow as of 2015, and interest in chatbots has grown. Thus, the start date is determined to 2014.

1.5. The Structure of Work

This work presents the incorporation of usability in chatbots development process and has been divided into the following chapters:

- The first chapter introduces the research work, both the approach to the problem and the possible solution and is the present chapter.
- The literature review regarding the defined research problem is presented in Chapter 2.
- Chapter 3 states the experimental setting includes methodology chosen. The experiment paid careful attention to control conditions to ensure a valid result.
- Chapters 4 shows the overall analysis which is according to the scheme presented in Chapter 3 and introduces novel aspects of the work.
- Chapter 5 brings the analysis of SOCIO.
- Finally, Chapter 6 details the conclusions obtained from the realization of this research work, making a synthesis and comparison of the cases of studies to discuss whether the incorporation of these conclusions was successful or not in chatbot SOCIO experiment, and describing future work.
- After the reference consulted and analyzed in the realization of this research, the appendices include primary studies (Appendix A), types of chatbots (Appendix B), experiment usability data (Appendix C), quality metrics (Appendix D), participant's preference statistics (Appendix E), task descriptions (Appendix F), questionnaires used in experiment (Appendix G) and ideal class diagram (Appendix H).

1.6. Contribution

For the Table 1.1 presents the contribution for the tasks carried out in this research. For the contribution, the status of the publication and the events are specified in Table 1.1. The state of the publication is published (P).

Table 1.1: Contribution derived from the research

Task	Contribution/Results	Type	State	Where
Review of Literature	Chatbot usability is a very incipient field of research, where the published studies are mainly surveys, usability tests, and rather informal experimental studies. Hence, it becomes necessary to perform more formal experiments to measure user experience, and exploit these results to provide usability-aware design guidelines.	Conference paper	P	SEKE2019

International Conference

- **Ranci Ren**, John W. Castro, Silvia T. Acuña and Juan de Lara (2019). Usability of Chatbots: A Systematic Mapping Study. In *Proceedings of the 31st International Conference on Software Engineering & Knowledge Engineering (SEKE 2019)*. Lisbon (Portugal), pp. 479-484 DOI: 10.18293/SEKE2019-029.
  Quality Index: **Core B**
  Relationship with Master's Dissertation: This work is described in Chapter 2.

CHAPTER 2

LITERATURE REVIEW

The literature review allows us to find and analyze the publications related to the areas in which we want to investigate, this is the first step to start a research project. This chapter presents a picture of the current state of the publications on usability of chatbots. For this, a review process known as the Systematic Mapping Study (SMS) has been conducted. According to [21] an SMS consists of a broad review of the relevant literature (primary research studies) in a specific thematic area, which aims to identify what available evidence exists on a topic [21]. We present an SMS to classify the applied usability techniques, the measured usability characteristics, the research methods used to evaluate the chatbots usability and the types of chatbots.

Chapter organization. In Sec. 2.1., we present related works. In Sec. 2.2., we describe the research method of the SMS. Sec. 2.3., presents the results of the SMS.

2.1. Related Works

According to the literature we reviewed, there are various of chatbots has been developed and evaluated usability. However, there are few works that discuss the usability of chatbots in an integrated and formalized manner. They evaluate the usability of chatbot through questionnaires, interviews, etc. For example, Jain et al., use questionnaire SUS to measure the usability of chatbot Convey [19] and Sinoo et al., scored the usability of the PAL chatbot on questionnaires with different Likert-scales [42].

Besides, when we reviewed the related literature, a set of papers in usability of conversational agents and computational dialogue systems also are informative to chatbot usability [38][40][9][20][50]. It is essential to make it clear how chatbots relate to its related glossaries. Conversational agents communicate with users in natural language (text, speech, or even both) that are already widely used commercially [20]. The conversational agent is recognized as a class of dialog systems [38], it generally fall into two classes: Embodied Conversational Agents and Chatbots. Therefore, chatbots are one category of conversational agents, which are software systems that mimic conversation with human users, but not typically humanoid robots [40]. The relations between these terms are shown in Figure 1. In this work, we only discuss chatbots.

We found only three systematic reviews related to chatbots [22][40][23]. The one by Klopfenstein et al. [22] surveys conversational interfaces, patterns, and paradigms. However, they do not detail the literature retrieval process, and hence may be potentially incomplete. The survey traces the history of chatbots, from ELIZA to modern chatbots for MOOCs. They conclude that only a subset of chatbots are designed for communicating in natural language, which sometimes makes users disappointed. They identify a category of conversational agents they call "*Botplications*". These are advanced agents following a set of simple and purposeful principles to provide access to services

and data. They have characteristics like history awareness, enhanced user-interface, limited use of NLP, message self-consistency and guided conversation. Then they compare features of major messaging platforms that support bots, like Messenger, WeChat, Line and Skype. Most of them already support a variety of message types, pictures, videos, and sounds. However, none of them have comprehensive enough features. For example, Line has groups, buttons and carousel features, but no payment and quick message reply. They detail advantages of bots for users and developers, and conclude stressing the benefits of chatbots as a new software platform to provide services and data to users.

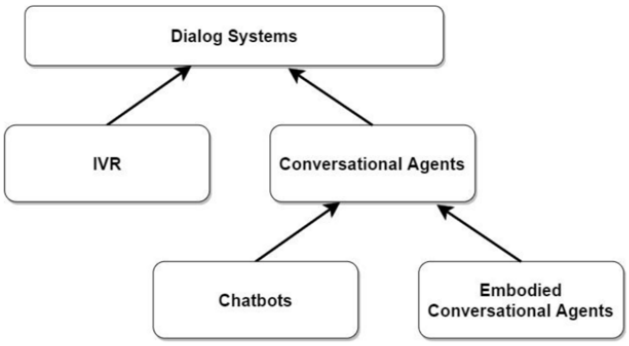

Figure 2.1: Relationships between classes of software-based dialog systems

The work by Ramesh et al. [40] surveys design techniques for conversational agents. The paper presents various solutions for building chatbots, including AIML, NLP and Natural Language Understanding (NLU). The authors describe the general structure of a chatbot, which consists of a responder, classifier and graph master. Then, they list several design techniques for chatbots, from pattern matching, to recurrent neural networks. They stress that NLP techniques are increasingly being used in recent years. The paper presents a classification of chatbots, which includes retrieval-based and generative-based, long and short conversations and open/closed domain. Retrieval-based chatbots pick responses from a pool of predefined ones. These bots do not make grammatical mistakes but do not have the capacity to mention earlier information in the conversation. Generative-based chatbots are able to refer to earlier information, but need more training and investigation. Long and short conversations are different in response to a single input or a long coherence input. Open and closed domains refer to whether the chatbot focusses on a specific task, or has open knowledge.

The third work, by Laranjo et al. [23] makes a systematic review of conversational agents in health. This review retrieved 1,513 research papers, and identified 17 primary studies. The search was performed in April 2017 and updated in February 2018. They describe 14 different conversational agents distinguishing type of communication technology, dialogue management, dialogue initiative, input modality and task-oriented aspects. The evaluation measures were divided into three main types: technical performance, user experience and health-related measures.

Overall, these works do not focus on usability characteristics or usability techniques of chatbots. Therefore, to the best of our knowledge, there is no SMS on the status of chatbot usability. Our work covers this gap.

2.2. Research Method

In this section, we aim to answer the following research questions:

RQ1: What is the state of the art of usability in the development of chatbots?

RQ2: How to evaluate the usability of chatbots using HCI principles?

To answer both questions, we have executed an SMS, to identify and classify these issues in the published literature [36]. The research method to find the literature are described in the following sections.

2.2.1. Search String Selection

The first step is identifying search strings and relevant keywords. For this purpose, several options were tried and the best one chosen. In particular, we first read some initial articles, obtaining keywords and basic knowledge related to the topic. By using different synonyms, four search strings were obtained (see the second column of Table 2.1).

- The first search string has two components. The first related to usability and user experience, the second related to chatbots.
- The second search string does not use synonyms for usability, because this word is the appropriate term in the area of Software Engineering to refer to the quality feature of the software concerning usability
- The third search string, unlike the previous ones, has three components. The first is related to usability. The second considers technique or practice. Finally, the third is related to chatbots.
- The last search string is the same as the previous one, but with a new synonym for technique and practice (i.e. evaluation). This synonym was not obtained from a particular paper, but included because it is likely that the literature refers to usability evaluation.

To choose the best search string it is necessary to test each of the previous search strings in each database. Table 2.1 presents the data returned by the search in each database.

After analyzing the data from each database and combining the opinions of two experts in HCI, we opted for the first string. This is so as the results are more balanced between the ACM Digital Library, SpringerLink and ScienceDirect databases (DBs). Moreover, it is the string that obtains more records from Scopus, which is the most complete and most used database. The final search string used in the SMS is shown in Table 2.2.

2.2.2. Databases and Search Protocol

Table 2.3 reports the search fields used for each DB. The search was performed in sequence from Scopus, ACM Digital Library, IEEE Xplore, SpringerLink and ScienceDirect. The search fields used were determined by the options provided by each DB.

Considering that the concept of chatbots is still relatively new, the search range is from January 2014 to October 2018. We ordered the search considering the DBs that returned most results. The search fields were selected to assure that searches were similar across DBs. The criteria used to retrieve the fundamental studies are summarized below.

Inclusion criteria:

- The paper is written in English; AND
- The abstract or title mentions an issue regarding chatbots and usability; OR
- The abstract mentions an issue related to usability engineering or HCI techniques; OR
- The abstract mentions an issue related to user experience.

Exclusion criteria:

- The paper does not present any issue related to chatbots and usability; OR
- The paper does not present any issue related to chatbots and user interaction; OR
- The paper does not present any issue related to chatbots and user experience.

Table 2.1: Results from each database

ID	Search strings	Number of papers found by search strings in				
		IEEE Xplore	***Scopus***	***ACM Digital Library***	***Springer Link***	***Science Direct***
1	(usability OR "usability technique" OR "usability practice" OR "user interaction" OR "user experience") AND (chatbots OR "chatbots development" OR "conversational agents" OR "chatterbot" OR "artificial conversational entity" OR "mobile chatbots")	5	105	20	21	19
2	(usability) AND (chatbots OR "chatbots development" OR "conversational agents" OR "chatterbot" OR "artificial conversational entity" OR "mobile chatbots")	5	72	10	32	16
3	(usability OR "user interaction") AND (technique OR practice) AND (chatbots OR "chatbots development" OR "conversational agents" OR "chatterbot" OR "artificial conversational entity" OR "mobile chatbots")	1	18	2	127	2
4	(usability OR "user interaction") AND (technique OR practice OR evaluation) AND (chatbots OR chatbots development" OR "conversational agents" OR "chatterbot" OR "artificial conversational entity" OR "mobile chatbots")	1	45	5	136	7

2.2.3. Paper Selection

The searches were run using the search strings and defined fields (Table 2.2 and Table 2.3). The number of papers returned by the first search was 170, which are called Retrieved Papers. Then by inspecting the title, keywords and abstract of each retrieved paper, 41 papers were filtered to the group of Candidate Papers. The whole group of

Candidate Papers was screened for duplicates. When duplicates were found, only the first occurrence of the paper was counted and maintained, the others were deleted. The final group has 39 papers, which is called Non-Duplicate Candidate Papers.

Each paper of the Non-Duplicate Candidate Papers group was read, to determine if they described any sort of usability of chatbots. The results were cross-checked by two experts in the HCI area, and any disagreement was discussed and resolved in our meetings. Finally, 19 papers were identified as primary studies, including a book chapter, conference papers, and journal articles. Table 2.4. summarizes the number of papers taken from each group, were all selected primary studies were finally taken from Scopus. Appendix A lists the primary studies located during the mapping study described in this section.

Table 2.2: Keywords used for the search string

Keywords		
"usability" **OR** "usability technique" **OR** "usability practice" **OR** "user interaction" **OR** "user experience"	**AND**	"chatbots" **OR** "chatbots development" **OR** "conversational agents" **OR** "chatterbot" **OR** "artificial conversational entity" **OR** "mobile chatbots"

Table 2.3: Search strings by DB

DBs	**Search Fields**
Scopus	"Title OR Abstract OR Keywords"
ACM Digital Library	"Abstract"
IEEE Xplore	"Abstract"
Springer Link	"Title OR Abstract OR Keywords"
ScienceDirect	"Title OR Abstract OR Keywords"

Table 2.4: The procedure of paper selection

DB	**Retrieved**	**Candidates**	**Non-Duplicate Candidates**	**Primary Studies**
Scopus	105	29	28	19
ACM Digital Library	20	5	4	0
IEEE Xplore	5	1	1	0
SpringerLink	21	6	6	0
ScienceDirect	19	0	0	0
TOTAL	**170**	**41**	**39**	**19**

2.3. Synthesis of the Results

Figure 2.2 provides an overview of the primary studies retrieved by the SMS. It is made of three categories, determined by the year of publication, type of paper (conference, journal, book chapter) and usability characteristics. The left-hand side is composed of two scatter (XY) charts with bubbles at the intersections of each category. The size of each bubble is determined by the number of primary studies that have been classified as belonging to the respective categories at the bubble coordinates. The right-hand side of Figure 2.2 indicates the number of primary studies by publication year. It can be seen that the number of publications started to grow from 2015, and many articles (mainly in

conferences) have been published each year since then, confirming the interest in the field. It can also be noted that most interest in chatbot usability is on effectiveness and satisfaction.

After conducting the SMS and analyzing the literature with respect to the usability of chatbots, the primary studies were classified from four different perspectives: usability techniques, usability characteristics, research methods and type of chatbots. These categories are reviewed next. The answer to RQ1 and RQ2 are in Section 6.1.

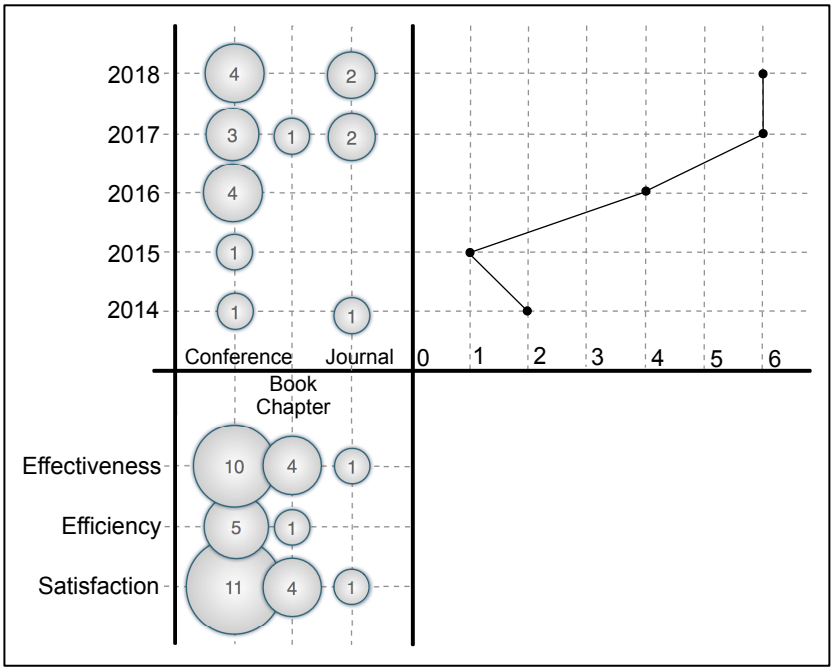

Figure 2.2: Overview of the primary studies

2.3.1. Usability Techniques

The primary studies in this category identify the usability techniques adopted from HCI. This is the second-most studied group in the literature. From the analysis of the papers, we found that questionnaires and interviews are most commonly used. The usability techniques are shown in Table 2.5.

In most cases, two or more techniques are combined for the usability evaluation. Each of these methods has its own characteristics, and cannot fully meet all the requirements of the usability test in isolation. Hence, it is necessary to combine various methods. For example, in [37], direct observation and the System Usability Scale (SUS) questionnaire are jointly used.

Table 2.5: Usability techniques

Usability Techniques	Primary studies
Questionnaire (SUS/ad-hoc)	[30][26][9][8][19][42][35][2][47][37] [10][41][48][11] [31][44]
Interview	[8][10][41][48][11][28][51][44][33]
Think-aloud	[8][47][41]
Direct observation	[37]
Cognitive walkthrough	[41]

In [10], questionnaires and interviews are used together in every research phase. In [41] the authors conduct a usability test to compare the usability of three chatbot platforms by using a SUS questionnaire, think-aloud and interview to rate the feedback from participants. A post-task questionnaire and an open-ended interview were used together in [44]. In [8], they video-recorded the whole experimental process for a retrospective think-aloud, and then conducted an interview and a questionnaire after accomplishing the tasks. In [11], user testing combines questionnaires and interviews. In [48], they conducted semi-structured interviews and used different standard questionnaires together in the first and the last period of evaluation. In the middle period, to gather more comprehensive information and attitudes of the users, they used the SUS questionnaire. In [26], they developed the pre-study questionnaire to illustrate the types of interactions that their participants perceived to be the most frequent with an Alexa chatbot. In [19], participants were asked to fill the metrics to measure their user experience, and also were asked to compare two different interfaces and justify their responses.

In some cases, the authors used just one single technique to measure usability. In [9], they conducted a survey using a questionnaire. In [42], though they mainly used questionnaires to measure usability, they track user's user experience through different questionnaires from different periods with open questions. In [28], they used structured interviews. In [35], different questionnaires were used in three different periods of the experiment. In [2], the authors used questionnaires to measure quantitative and qualitative evaluations of the new NLP method used by the chatbot. In [33], during the interview, participants had to explain the difficulties they had. In [51] to avoid excessive verbosity and to use verbal instead of text feedback, they used interviews with open-ended questions. In [31], to related to reliability, usability, and functionality of the system.

Overall, we can conclude that the technique used depends on the specific conditions, while there is no standard proposal.

2.3.2. Usability Characteristics

According to the primary studies, usability characteristics are mainly identified in three aspects: Effectiveness, Efficiency and Satisfaction.

1) *Effectiveness:* Effectiveness is defined as the accuracy and completeness with which users achieve specified goals in HCI [18] [16]. From Figure 2.2, most papers consider effectiveness as an essential factor when evaluating the usability of chatbots. Table 2.6 shows more details on the used effectiveness criteria. In particular, we have identified task completion, accuracy of chatbot reply, comparison with recall and expert assessment as the main means to assess effectiveness. In [9] by gathering feedback from experts and potential users, they evaluate grading of the perceived quality of effectiveness of the chatbot [16] and find some shortcomings and possible solutions that will enhance the application's usability for its intended audience.

In these works, the number of correct responses or interventions indicates the accuracy (to measure if users achieve specified goals [18]) and recall (users' ability to recall information from the interface [16]). The result shows that most chatbots achieve the required accuracy and recall of response [31][44]. For example, through comparing with other chatbots with similar functionality for completing the task, the authors in [19] proved their e-commerce chatbot performs better than the default chatbot. In [2], according to the result of the questionnaire, 80% participants claimed that the content of the retrieved information is clear and useful. In [28], the authors measure the number of users who complete the task (interview) through two different digital tools, showing that

the chatbot has higher acceptability. To identify the measures of characteristics accuracy and recall, the works [18] [16] have been followed.

Table 2.6: Effectiveness

Measures of Effectiveness	**Primary studies**
Task completion	[19][26][41][31][28]
Accuracy	[35][2][11][31][44][51]
Recall	[2][31]
Experts and Users' assessment	[9][8][10][11]

However, there are still some problems to meet the high level of task completion and accuracy of the chatbot reply. In [26], 19 incomplete tasks were reported among the participants. The reason why the task could not be completed is not due to the user, it's because of the system design. In [35], during the evaluation, there were some problems with the DBpedia semantic entry point, which affected the accuracy of some of the users. In [51], 46 entries were negotiated, of which 7 (15.2 %) did not correspond correctly to the user's original wishes, but when a participant used more lengthy sentences to express, he produced noticeably more utterances compared to the average of the others. This problem mainly resulted from the inability of the system to process long, convoluted utterances properly and it lacks the ability to guide the user during the interaction. In [11], the chatbot generated unnecessary information in response to highly structured conversations. In [31], the factor affecting the accuracy of chatbot reply is the need to handle one or more user conversation turns before providing the answer.

2) *Efficiency:* Efficiency relates to the resources expended in relation to the accuracy and completeness with which the users achieve their goals [18][16]. Most papers discuss task completion time, mental effort and communication effort to use the chatbot, as shown in Table 2.7.

Table 2.7: Efficiency

Measures of Efficiency	**Primary studies**
Task completion time	[19][33]
Mental effort	[19][30]
Communication effort	[35][37][10]

In [19], the authors compare the number of views and average time the participants took in completing a task with the Convey chatbot, and a default one. The results showed they spent more effort and time performing the task with Convey.

Perceived autonomy and competence are factors favoring efficiency in chatbot usability [30]. In [35], it was noted that, since the chatbot can correct erroneous inputs, users do not need to spend much communication effort when talking to the chatbot. In addition, less communication effort makes the chatbot easier to operate. In [37] the authors count the number of participants' cumulative assertions to measure the communication effort,

the steady increase demonstrates that users can use the chatbot efficiently in short time. Finally, users spend more communication effort when the chatbot has limited conversational ability, as discussed in [10].

3) Satisfaction: This is the largest group of papers within the primary studies. Satisfaction is defined as the degree to which user needs are satisfied when a product or system is used in a specified context of use [18][16]. The measures of satisfaction include ease-of-use, context-dependent questions, satisfaction before and during use, complexity control, physical discomfort of the interface, pleasure, the willingness of using the chatbot again, and enjoyment and learnability. From Table 2.8, the ease-of-use, willingness to use the chatbot again and user experience are the main measures of satisfaction used. Emotional aspects such as perceived utility, pleasure, comfort, are also considered in [11], and are related to the user experience. Among the primary studies, works have been found measuring the user experience mainly considering the physical discomfort and pleasure. These works are highlighted with a rectangle in the Table 2.8.

Table 2.8: Satisfaction

Measures of Satisfaction	**Primary studies**
Ease-of-use	[19][9][42][35][2][47][41][48][11][44]
Context-dependent question	[41][48][33]
Before use	[9][10][48]
During use	[30][10]
Complexity control	[9][2][48][31]
Physical discomfort	[9][2][10]
Pleasure	[19][26][42][35][10][41][44][33]
Want to use again	[19][9] [42][35][10]
Learnability	[8][42][41]

Frequently, chatbots can satisfy users simple needs, such as answering questions [26], buying goods [10], and answering simple questions in natural language [2][44]. But when users' needs become more complex, involving emotional needs and intensive interactions, the satisfaction with chatbots typically declines. On the one hand, chatbots lack personality. For example, a digital pet was proposed [10] to accompany the elderly, but some users commented that the relationship between users and chatbots was superficial because of the system's limited conversational ability and its occasional one-way communication pattern: "[The digital pet] can't tell me anything about his personal life".

On the other hand, chatbots have more exploration space for interaction with users. A physical chatbot was proposed in [42] to support self-management of diabetes by children. The usability evaluation included capabilities, social presence, and the quantity of speech and movements. Children stated that the physical chatbots were more (inter)active, more present and capable of doing different things, such as dancing. Chatbots with actual images or entities are more likely to establish relationships with users, improving their experience. In [33], a combination of speech-and-gesture makes users get along better with the chatbot. In [41], the authors compared Pandorabot with two other chatbots. Overall, Pandorabot's voice sounded less robotic, entertaining users better. In [41] some participants claimed that they have been helped the most by the agent's voice, improving their domain knowledge. It seems that the use of speech synthesis constitutes a crucial

design factor, even for text-based dialogue systems. Chatbots with actual images, entities or voice are more likely to establish relationships with users, thus improving users' experience.

Besides, more flexibility and speech command context-dependence are required for better usability. In [19] participants mentioned that a shopping chatbot was easy to use since it tracked their search history. In [35] some users do not consider they need an affective enhanced semantic chatbot at home. In [48] the authors observed that the acceptance of the chatbot decreases since its response mismatched the users' initial expectations. Potential explanations for such inconsistencies might include fundamental differences in user expectations for the chatbot and the emphasis on the interactive and entertaining qualities of the system over its informational value.

The user background should be considered a key point in evaluating satisfaction. Cultural, socio-economical and personal preferences can influence the opinions towards chatbots. In [48], the authors noticed that users in the Netherlands were more experienced with technology than in the other two countries of the study, therefore their expectations towards the novel technology were higher. In [8] users with higher technical knowledge learned quicker to use the chatbot.

2.3.3. Research Methods

The research methods used by the authors of the primary studies within this group include surveys of chatbots users' experience, experiments of using chatbots to realize some given tasks, usability tests, case studies and quasi-experiments. The research methods are detailed in Table 2.9. The number of papers using different methods is shown in Figure 2.3. The most common research methods include survey, experiment and usability tests.

In most experiments, very simple tasks are proposed. For example, using Apple Siri to find an inexpensive hotel in Osaka [8], search a flight ticket and hotel room via the chatbot [30], whether a simple chatbot can be appropriately used as a delivery mechanism [11], buying shoes [19], measuring the quality and quantity of the information retrieved [2], taking a structured interview with chatbot [28], or playing a game [33]. However, real-life situations are more complicated.

Table 2.9: Research methods

Research methods	**Primary studies**
Survey	[26][42][47][41][11][31][44]
Experiment	[2][30][26][9][8][42][35][47]
Usability test	[2][47][41][31][28][51]
Case study	[9][37][48][33]
Quasi-experiment	[37][10][31]

Rather than aiming to fully recreate the real-world task, simulation-based assessment should incorporate psychologically relevant aspects and situations from the real-world task and environment, such as time-pressure, or high uncertainty. In [35], the authors show that when the chatbot has visual appearance and emotions, users do not notice the small change of voice and facial expression. The experiment concludes that it is not necessary to use extremely accurate facial expressions for realistic use. In [10], the authors deployed a digital pet avatar in the participants' home for 3 months to simulate real-life situations as much as possible. In [37], experiments were designed in a complex

way, to simulate real-word situations. This is sometimes necessary, because if we were unable to use the chatbot effectively with a design as realistic as possible (but nonetheless simplified), it would be unlikely to be effective under more challenging conditions for the military, law enforcement and others in safety-critical real-world environments.

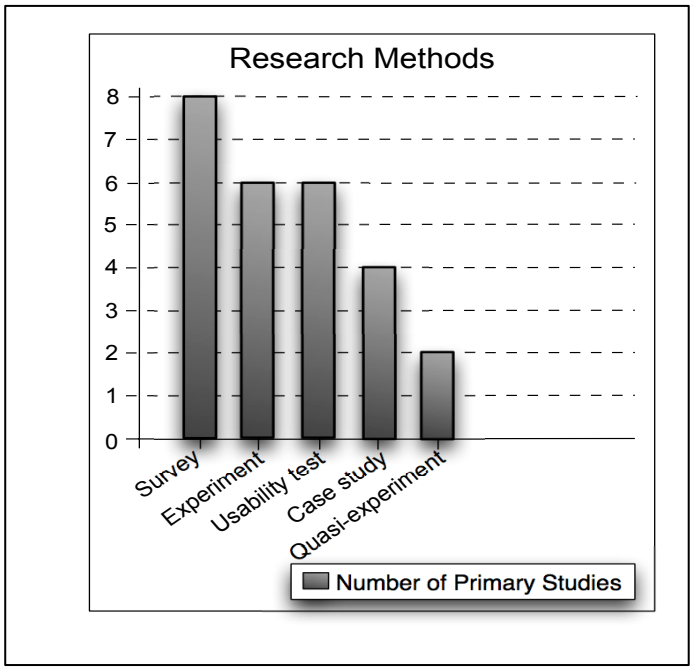

Figure 2.3: Number of primary studies by research methods

In most cases, researchers make comparisons. For example, in [9], the authors compare a chatbot with a similar one or with a similar application. This way, he can better understand user satisfaction. In [48], the authors conducted the case study in three countries, which proves that different culture and background has influence in the evaluation. Research methods can also be combined, which typically yields better results. In [26], a pre-survey questionnaire was performed to assess the usability of an Alexa chatbot. Then, an experiment was conducted to investigate specific problems.

Machine learning (ML) algorithms, in combination with cloud-based databases can be used to solve some current shortcomings of handling natural language. For example, chatbots can't recognize the words that they haven't been programmed for, and some chatbots speak unnatural language [11][44]. ML and related NLP technology help to bring more accurate and flexible responses to users, improving users' confidence and comfort in using the chatbot [41].

2.3.4. Type of Chatbots

The AIML technology is still widely used in the design of chatbots [11][31]. However, the use of chatbots using NLP is growing [37]. For example, the PAL project [42] can generate reasonable feedback through user-entered information. In [26], the authors show that the chatbot can be used via natural language phonic control, to perform searches, entertainment, and to control other devices. In [9], more users are satisfied with the chatbot, due to its speaker functionality and natural conversation flow. In [2], the authors used Object Relational Mapping (ORM) frameworks to improve the process of generating SQL statements from NL queries.

Many chatbots are built as Embodied Conversational Agents (ECA), and there are increasing number of chatbots with image, sound and personality [35][10][48][33]. However, sometimes chatbots have negative emotions. When the ECA has a negative personality, it tends to ignore or blame the user [35]. In addition, the chatbot is required to have the ability to learn and adapt to its user context to be useful [48]. Therefore, complete evaluations should be carried out to obtain a better comprehension of these issues. In this respect, [19] built an e-commerce chatbot using IBM's Watson assistant, and compared with a default chatbot. The authors concluded the new chatbot brought users more effective experience since it takes less mental and physical effort to obtain the product.

Based on the set of papers I read, I analyze and summarize the types of chatbots with nine aspects of investigation: year, program name, designer name, description of functions, design techniques, user experience, usability techniques, usability characteristics, and research methods, as shown in Appendix B.

CHAPTER 3

EXPERIMENTAL SETTING

This chapter describes the method used to conduct the experiment. In this chapter, we present our goal, research questions and hypotheses firstly. Second, we perspectively indicate the overall description of the experiment, participants' choices and detailed execution of the experiment. Then we bring the criteria and metrics of evaluating experiment. Finally, we explain the sheet we used to collect and calculate data. To obtain an effectiveness research result, we conducted a control experiment with 2 groups: A group that will use a chatbot to make a class diagram of an application, and the other group that will make the class diagram in Telegram platform. Based on the statement of the research question to be addressed together with the research hypotheses, we applied an experimental design that evolved into the final design. In the following, we discuss the experimental design and the process enacted.

3.1. Goal, Research Questions and Hypotheses

The **objective** of the research is to evaluate and compare **usability** with respect to efficiency, effectiveness and satisfaction, as well as to compare the **quality** of the class diagrams obtained.

In accordance with this objective, the research question is as follows:

> **RQ3:** Does the use of the chatbot SOCIO has a positive effect on the efficiency, effectiveness and satisfaction of the participant when making a class diagram, as well as its quality?

The hypotheses for both SOCIO and CREATELY are the following:

> **H.1.0** There is no significant difference in efficiency with SOCIO or with CREATELY when making the class diagram.
>
> **H.2.0** There is no significant difference in effectiveness with SOCIO or with CREATELY when making the class diagram.
>
> **H.3.0** There is no significant difference in satisfaction with SOCIO or with CREATELY when making the class diagram.
>
> **H.4.0** There is no significant difference in the quality of the class diagram performed using SOCIO or CREATELY.

The hypotheses only for SOCIO are the following, each variable used to evaluate the efficiency of SOCIO is considered as a metric:

Metric 1: Number of all messages sent to chatbot SOCIO

H.S.1.0 (null). There is no significant difference in the number of all messages sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

H.S.1.1 (alternate). There is a significant difference in the number of all messages sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

Metric 2: Number of error messages sent to chatbot SOCIO

H.S.2.0 (null). There is no significant difference in the number of error messages sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

H.S.2.1 (alternate). There is a significant difference in the number of error messages sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

Metric 3: Number of useful messages sent to chatbot SOCIO

H.S.3.0 (null). There is no significant difference in the number of useful messages sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

H.S.3.1 (alternate). There is a significant difference in the number of useful messages sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

Metric 4: Number of descriptive messages sent to chatbot SOCIO

H.S.4.0 (null). There is no significant difference in the number of descriptive messages sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

H.S.4.1 (alternate). There is a significant difference in the number of descriptive messages sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

Metric 5: Number of command sent to chatbot SOCIO

H.S.5.0 (null). There is no significant difference in the number of commands sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

H.S.5.1 (alternate). There is a significant difference in the number of commands sent to chatbot SOCIO by the teams when performing Task 1 or Task 2.

Metric 6: Number of actions triggered by chatbot SOCIO

H.S.6.0 (null). There is no significant difference in the number of actions triggered by chatbot SOCIO when teams perform Task 1 or Task 2.

H.S.6.1 (alternate). There is a significant difference in the number of actions triggered by chatbot SOCIO when teams perform Task 1 or Task 2.

3.2. Description

In the work, two tools **SOCIO** and **CREATELY** be used to execute the experiment. SOCIO is a chatbot that helps in creating the class diagrams by interacting in natural language over Telegram. **CREATELY** (creately.com) is a tool for creating diagrams including class diagrams.

Table 3.1: Experimental design

Tool	Task Period Sequence	Task 1 Period 1		Task 2 Period 2	
		SC	CR	SC	CR
Group 1: SC-CR		X	—	—	X
Group 2: CR-SC		—	X	X	—

The experiment presents a **within-subjects cross-over design** of 2 **sequences** and 2 **periods** (see Table 3.1). The participants are divided into two groups **(Group 1 and Group 2)** randomly. Group 1 is associated with the SOCIO-CREATELY sequence (SC-CR) and Group 2 is associated with the CREATELY-SOCIO sequence (CR-SC). Both groups, in teams of 3 (formed at random), will perform two tasks (**task 1 and task 2**), in each of the task a class diagram will be made. The tasks are associated to the periods, and each task is done with a different tool, depending on the sequence associated with the group. In any case, participants in the same group are not be allowed to talk with each other, all conversations and communications were taken place in the Telegram group to ensure we can fully record the experimental data.

3.3. Participants

The experiment was carried out by a total of **54 participants**, which have a degree in Computer Science or related degree at the *Universidad de la Fuerzas Armadas ESPE Extensión Latacunga* in Ecuador.

All of them have studied or are studying the subjects of Software Analysis and Design and Software Analysis and Design project, thus they have the necessary knowledge to make a class diagram.

3.4. Execution of the Experiment

The experiment was carried out over 4 sessions. The participants were assigned to each of session according to their time schedule. **Sessions 1 and 2** constitute **Group 1**, and **sessions 3 and 4** constitute **Group 2**. The determination of the groups is random. Once each group of 27 participants was formed, they were divided into 9 teams randomly, thus obtaining a total of **18 teams**. First 4 teams of Group1 executed the tasks in Session 1, In Session 2, teams 5-9 of Group 1 performed the tasks. Team 10-14 which are first 5 teams of Group 2 are performed the tasks in Session 3. Team 15-18 performed the tasks in Session 4. (The arrangement is shown in Table 3.2.)

Table 3.2: Sessions, groups and teams

	Group	Teams
Session 1	1	1/2/3/4
Session 2		5/6/7/8/9
Session 3	2	10/11/12/13/14
Session 4		15/16/17/18

In each session, we had not prepared any preparatory or practice session with the subjects. Firstly, we asked that each of participants sign an **informed consent** stating that they allowed us permission to record their sessions via Telegram. Then, each participant was required to complete a **familiarity questionnaire** (see Appendix G) purposely designed to help us collect their basic information such as their age, their gender, their level of English, also let us further explore each participant's preconceived ideas regarding topics in which they were asked for their use of social medias and their level of knowledge about class diagrams.

After that, according to the sequence to which they were associated, participants received a brief **tutorial** of the tool they had to use in the **first task**. Then, they were required to perform the first task in a maximum of 30 minutes. The first task is consisted in designing the class diagram of a **store** that wanted to manage their products and their customers, and at the end they filled out a modified **satisfaction questionnaire System Usability Scale (SUS)** (see Appendix G) associated with the tool they used in the first task. The SUS questions were asked with a rating of 1 to 5 with 1 representing "strongly disagree" and 5 representing "strongly agree".

Once the questionnaire was completed, they received a **tutorial** of the tool they had to use in the second task in the same way depending on the sequence to which they were associated. Then, they performed the **second task** in a maximum of 30 minutes which consisted in designing the class diagram of a **school** that wanted to manage their subjects and their students. At the end they filled in another modified **satisfaction questionnaire SUS** associated with the tool they used in the second task. The rating criteria are same as before. But, in this last questionnaire, they were specifically asked they preferred SOCIO or CREATELY. Figure 3.1 shows the detail of each session.

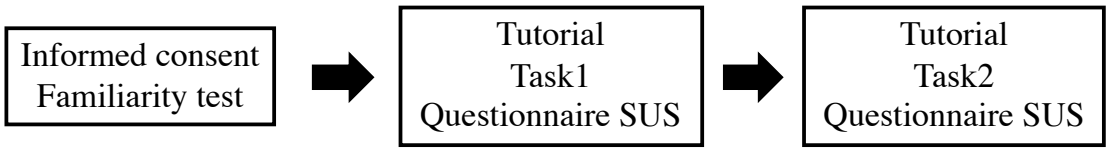

Figure 3.1: Experimental procedure

When performing tasks with chatbot SOCIO (@SOCIO), it is necessary to have a Telegram group in which the chatbot is a member. To communicate with SOCIO, participants have to send messages as commands to SOCIO, once SOCIO receives them, it interprets them and sends an answer to participants. Within those messages, descriptive messages (e.g., "/ talk The house contains doors") or imperatives (e.g., "/ talk Add house, add doors") can be used to create the diagram. From these messages, SOCIO modifies the diagram and sends an image with the changes. This Telegram group is also used for the members to discuss and to make the appropriate decisions.

During the completion of the tasks, if the team carried out the task with SOCIO, the members used a Telegram group, adding the chatbot as a member, to communicate with each other and with SOCIO. If the team carried out the task with CREATELY, they also used a Telegram group, through which they had to communicate in order to organize themselves and make decisions.

3.5. Facts, Various Answers and Metric

With regard to the **response variable**, ISO/IEC 25010 [18] defines measures of what the standard refers to as quality in use; and efficiency, effectiveness and satisfaction are common attributes for evaluating product usability. Besides, we used **quality** to measure the quality of experiment results (class diagrams obtained). Thus, the **response variables** of our experiment are **efficiency**, **effectiveness**, **satisfaction** and the **quality of class diagram.** The response variables and their respective **metrics** are outlined below.

The metrics that have an asterisk will be used for the usability of both tools, those that do not have it are metrics only related to SOCIO.

The **metrics** used to measure **efficiency** are:

- **Speed**:
  - **Time** measured in minutes taken by a team to complete the task, with a maximum of 30 minutes (*).

- **Fluency**:
  - **Number of discussion messages** generated by a team during the completion of the task (*).
  - **Number of all messages send to the chatbot** (includes error messages without and with input after the command / talk).
  - **Number of invalid messages addressed to the chatbot** (includes messages have intention of sending to SOCIO, but failed due to writing errors or in the case of being sent, but failed to be understood by SOCIO).

- **Interactivity**:
  - **Number of messages addressed to the chatbot** that contribute to create the diagram by a team. From these messages, number of commands and number of descriptive messages.
  - **Number of actions** triggered by those messages that have contributed to create the diagram by a team.

The **metric** used to measure the **effectiveness** is **completeness**, based on the perceived success in carrying out the task. That is, if the task has been carried out in its entirety (*).

Satisfaction will be measured with the mean value of the responses to the SUS questionnaire questions and sentiment analysis of 3-4 open-ended questions. The values of the questionnaire responses are ordinal on a 5- point Likert scale rating between 1 (disagree completely) and 5 (agree completely) (*).

To explore and visualize the data, we used the R language for sentiment analysis to the answer of open-ended questions and counted the word frequency within the answer. The NRC Emotion Lexicon [29] is conducted by NRC-Canada sentiment analysis system, it

is a list of words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).

We used emotional dictionary NRC with the line of code “nrc<- get_sentiment_dictionary ('nrc', language = "spanish")”. However, if there is no word in the data that corresponds to a certain emotion, then there will be no such emotion in the final result. Each response is treated as a unit in the chart, and each unit has a different mood, and each emotion is collected separately and displayed in a histogram.

Among the ten emotions, negative and positive are treated as two broad categories. Negative contains anger, disgust, fear, and sadness. Positive contains anticipation, joy, surprise and trust. Some words have negative or positive emotions but are not specifically expressed in 8 subcategories, which are counted as negative or positive emotions.

In addition, each word may contain more than one emotion, such as “recommendable” for positive, joy, and trust. The emotions reflected by all the words in each answer are divided into 10 categories, and the weight of each emotion in each answer is the same. Each response is treated as a unit in the chart displayed in the histogram.

Before we did the analysis, we sorted the word frequency and follow it to clean the stop words, such as “trabajo”, it doesn’t has certain emotional direction in our case, but somehow in other situation it is treated as a positive word, so we have to clean the stop words manually.

We first combined all the response of questions to conduct an overall analysis. Then we analyze each question one by one.

As for **quality**, we analyzed the data of the experiment with Linear Mixed Models following Vegas et al. [49]. The factor addressed by the study is **three** identical factors which are common in the all linear mixed models: **(1) sequence** (either Group 1 or Group 2), accounting for the assignment of teams to a combination of task and treatment; **(2) order** (either Group 1 or Group 2), confounded with task, accounting for the task that the teams had to implement; and **(3) treatment** (either CREATELY or SOCIO), accounting for the tool applied by the teams to solve the tasks.

We complement the results of the statistical analysis with Cohen’s d for the treatments (d, hereinafter) and their corresponding standard errors (SEs). For this, we follow the formulae provided in the Cochrane Handbook for cross-over designs [15]. In the following we go over the analyses on the dependent variables.

The metrics used to mediate **quality** are:

- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **Accuracy** = TN + TP / (TP + FP + FN)
- **Error** = (FP + FN) / (TP + FP + FN)
- **Success** = TP / (#predicted diagram elements)

In the formulas above: true positives (TP), false positives (FP), false negatives (FN) are obtained by comparing the predicted class diagram and the objective class diagram.

In our case, actual corresponds to the predicted class diagram and the objective class diagram with each of the class diagram:

- TP (true positive): Number of the elements that are found in both the predicted class diagram and the objective class diagram.
- FN (false negative): Number of the elements that are found in the predicted class diagram, but not in the objective class diagram.
- FP (false positive): Number of the elements that are found in the objective class diagram, but not in the predicted class diagram.
- TN (true negative): In the comparison of models there is no true negative, the value is always 0.

The predicted class diagram refers to the class diagrams created by each team, the objective class diagram refers to the ideal class diagram (see Appendix H) of each task, performed prior to execution of the experiment and were recognized by experts in Software Engineering field. Elements refers to classes, attributes within each classes and relations between each two classes. They all have same weight.

3.6. Data Obtained

The *Usability Data* sheet (Appendix C) shows the efficiency, effectiveness and satisfaction data collected from the experiment. The *Quality Metrics* sheet (Appendix D) presents quality data of each team corresponding to the metrics shown in the Section 3.5. in the same order. The *Preference* sheet (Appendix E), perspective by group and by participant, indicates answers to the question of which tool they prefer.

CHAPTER 4

OVERALL ANALYSIS

This chapter describes the process of analyzing the experiment data comparing SOCIO and CREATELY. We conducted a cross-over experiment which avoids bias due to sequence of tasks or participants. Firstly, we analyzed the result of the familiarity questionnaire, then we compared SOCIO and CREATELY in the aspects of efficiency, effectiveness, satisfaction and quality.

4.1. Familiarity Questionnaire

According to the familiarity questionnaire, we collected basic information about our participants. Note the following details:

- The final sample consists of 54 subjects. Of the sample, 44 are men and 10 are women.
- Subjects have a mean age of 22 and a standard deviation of 1.74. The highest concentration of participants is in the range 21-23 years.
- 66.7% of subjects use social media frequently. WhatsApp, Facebook, Instagram and Telegram are the most used social medias by participants.
- All participants believe they have knowledge about class diagrams. 90% of them relatively familiar with class diagrams.
- 87.1% of the participants have used the Telegram application or use it frequently, while 12.9% have no experience in using it.
- In relation to chatbots, all participants consider they understand them at least at the conceptual level. Regarding their usage habits, 29.6 % have never used a chatbot, while 70.4% have experience (55.6% have used chatbots at times and 14.8% are regular users). The fact of having subjects lacking previous experience of using chatbot is a beneficial factor for the experiment, as it contributed to the greater sensitivity to the usability of the tool and the validity of our results.
- Although no subject is a native English speaker, all of them consider having a fluent level of English.

4.2. Efficiency Analysis

4.2.1. Speed

Figure 4.1 shows the box-plot corresponding to the time spent by the teams per treatment. As we can see in Figure 4.1, time spent seems to be less on SOCIO than CREATELY. Table 4.1 shows the results of the linear mixed model we fitted to analyze the data. As we can see in Table 4.1, neither the sequence nor the order has a statistically significant impact on time spent, only the treatment. Finally, d=0.80, SE(d)=0.41, suggesting that a large effect size -according to rules of thumb [4]- materialized for the treatment in terms of time spent. In sum, **SOCIO saved more time than CREATELY**.

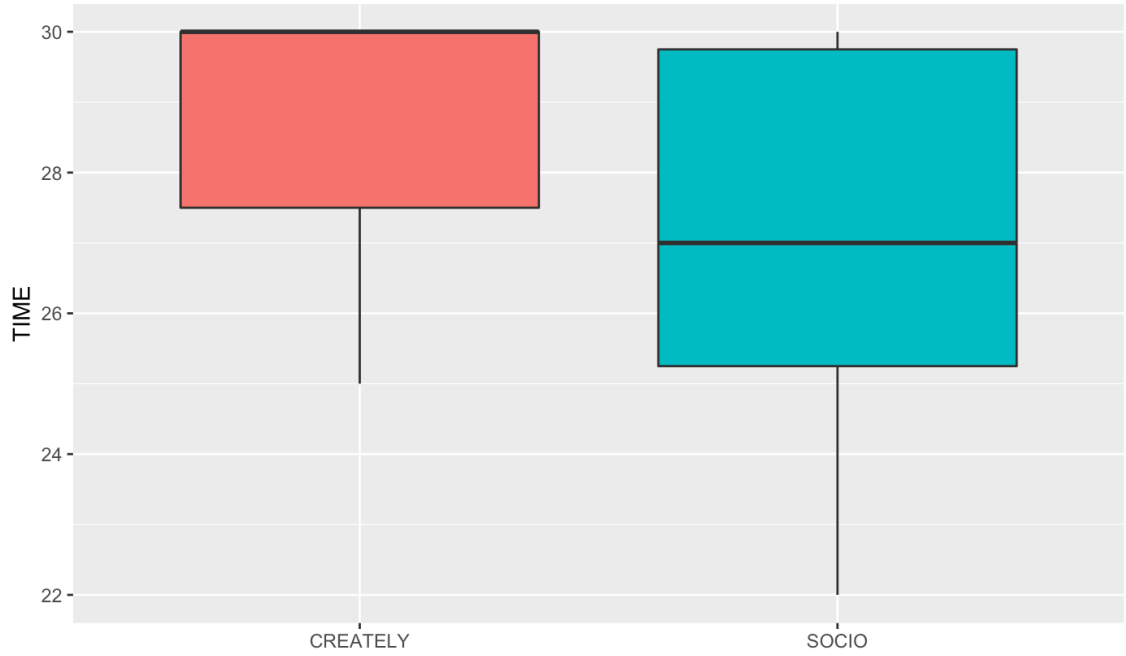

Figure 4.1: Time spent on completing the task by CREATELY and SOCIO

Table 4.1: Linear Mixed Model for time

	Estimate	**Std. Error**	***p*-value**
(Intercept)	27.89	0.73	0.00
Seq	1.11	0.73	0.15
Treatment	-1.78	0.73	**0.03**
Order	0.78	0.73	0.30

4.2.2. Fluency

Figure 4.2 shows the box-plot corresponding to the number of discussion messages by the teams per treatment. As we can see in Figure 4.2, the number of discussion messages seems to be less for SOCIO than CREATELY. Table 4.2 shows the results of the linear mixed model we fitted to analyze the data. As we can see in Table 4.2, only the treatment has a statistically significant impact on number of discussion messages. Not the sequence nor the order. This indicated that **SOCIO saved more communication than CREATELY in terms of number of discussion messages**. Finally, d= 0.70, SE(d)=0.22, suggesting that a relatively large effect size -according to rules of thumb [4]- materialized for the treatment in terms of completeness.

All in all, efficiency is measured by the amount of time it takes to finish the tasks and the number of discussion messages. In both aspects, **working with SOCIO seems more efficient and it reduced communication costs.**

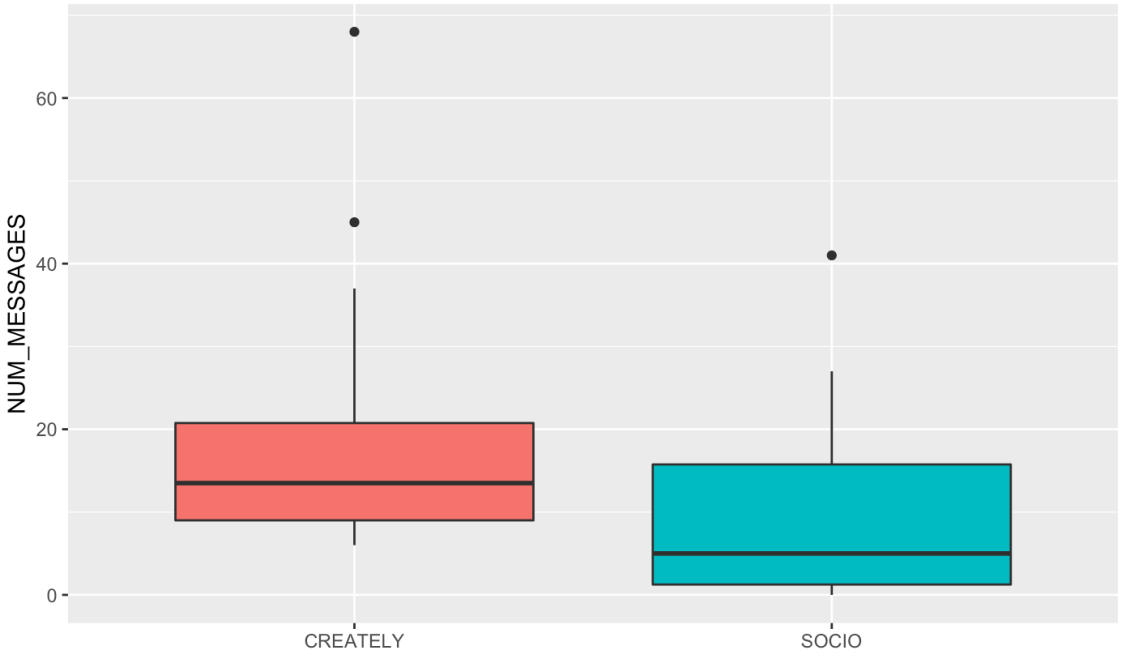

Figure 4.2: Number of discussion messages for CREATELY and SOCIO

Table 4.2: Linear Mixed Model for number of discussion messages

	Estimate	**Std. Error**	***p*-value**
(Intercept)	22.72	4.78	0.0002
Seq	-3.17	6.10	0.61
Treatment	-9.94	2.92	**0.0036**
Order	-3.17	2.92	0.29

4.3. Effectiveness Analysis

4.3.1. Completeness

Figure 4.3 shows the box-plot corresponding to the completeness scores of the teams per treatment. As we can see in Figure 4.3, completeness scores seem higher and less sparse for SOCIO than CREATELY.

Table 4.3 shows the results of the linear mixed model we fitted to analyze the data. As we can see in Table 4.3, only the treatment has a statistically significant impact on completeness at 0.006 level, not the sequence nor the order. Finally, d=-1.05, SE(d)=0.41, suggesting that a very large effect size -according to rules of thumb [4]- materialized for the treatment in terms of completeness.

This implies that **SOCIO outperformed CREATELY in terms of completeness**.

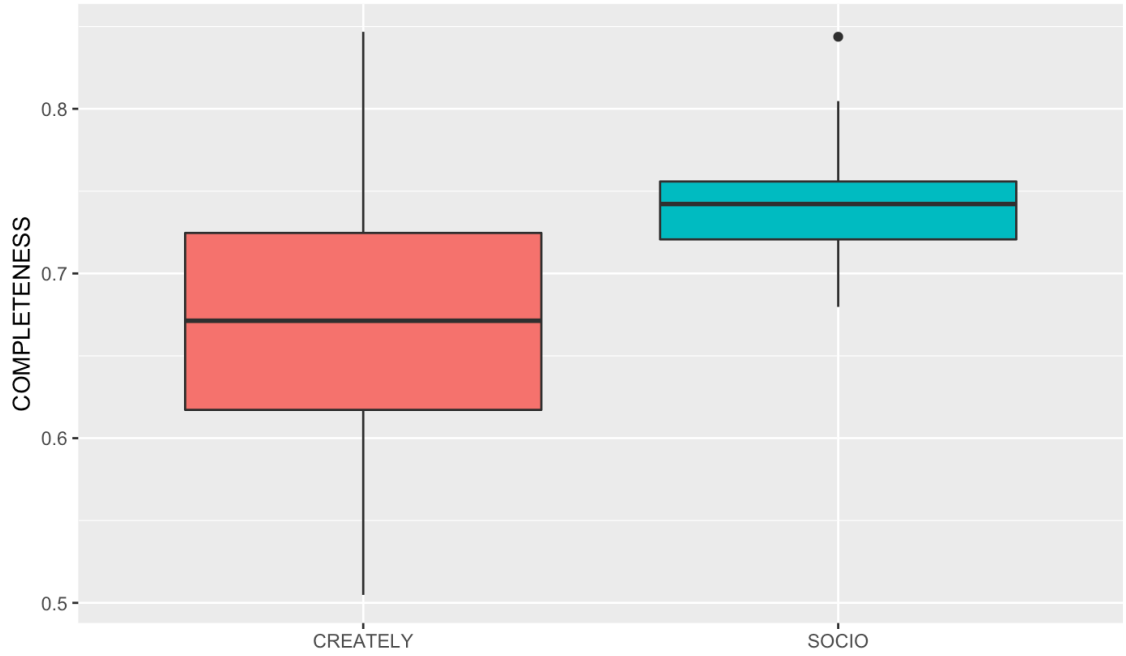

Figure 4.3: Completeness scores for CREATELY and SOCIO

Table 4.3: Linear Mixed Model for completeness

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.68	0.02	0
Seq	0.0004	0.02	0.99
Treatment	0.07	0.02	**0.006**
Order	-0.016	0.02	0.496

4.4. Satisfaction Analyze

We used questionnaire to evaluate satisfaction towards two tools. Each questionnaire includes 10 questions of SUS ,which are answered on a 5-point Likert scale with 3-4 additional open questions at the end.

4.4.1. Open-ended Questions

From Figure 4.4 and Figure 4.5, it can be seen that SOCIO has apparently more positive responses than CREATELY, it pleased the participants. However, it also received more negative comments. Both of them gain relatively trust by the participants.

Q: Please indicate three positive aspects that you want to highlight about the tool.

As the Figure 4.6 and Figure 4.7, SOCIO obtained many more positive comments than CREATELY, in aspects of anticipation, joy and surprise as well. The filtered (mentioned more than 3 times) positive aspects of both tools which satisfied participants are shown in Table 4.4 and Table 4.5. It can be seen that SOCIO has more positive comments and aspects. Both tools satisfied participants by quick responsiveness, easy of use and collaborative work, but for SOCIO they expressed satisfaction with more objects than for CREATELY in these three aspects. Besides, CREATELY was praised for its friendly

interface. Some claimed that the use of a chatbot made the technology more user-friendly by allowing the user to have a more entertaining interaction. In other words, SOCIO was better suited to entertain the user.

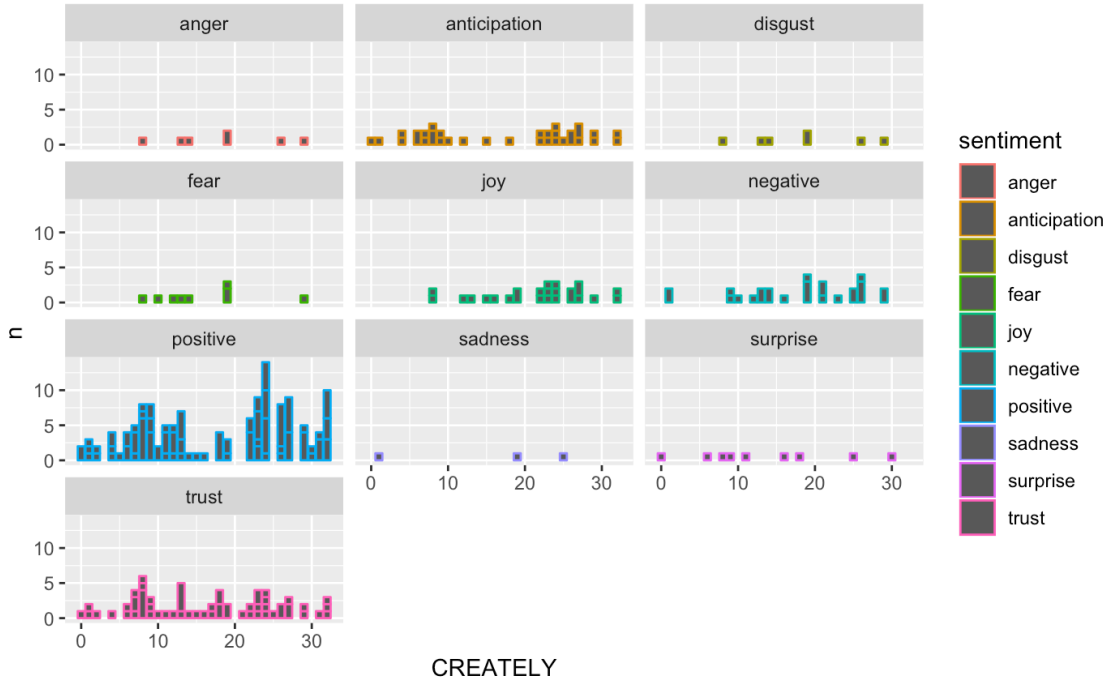

Figure 4.4: Overall satisfaction analysis of open-ended questions for CREATELY

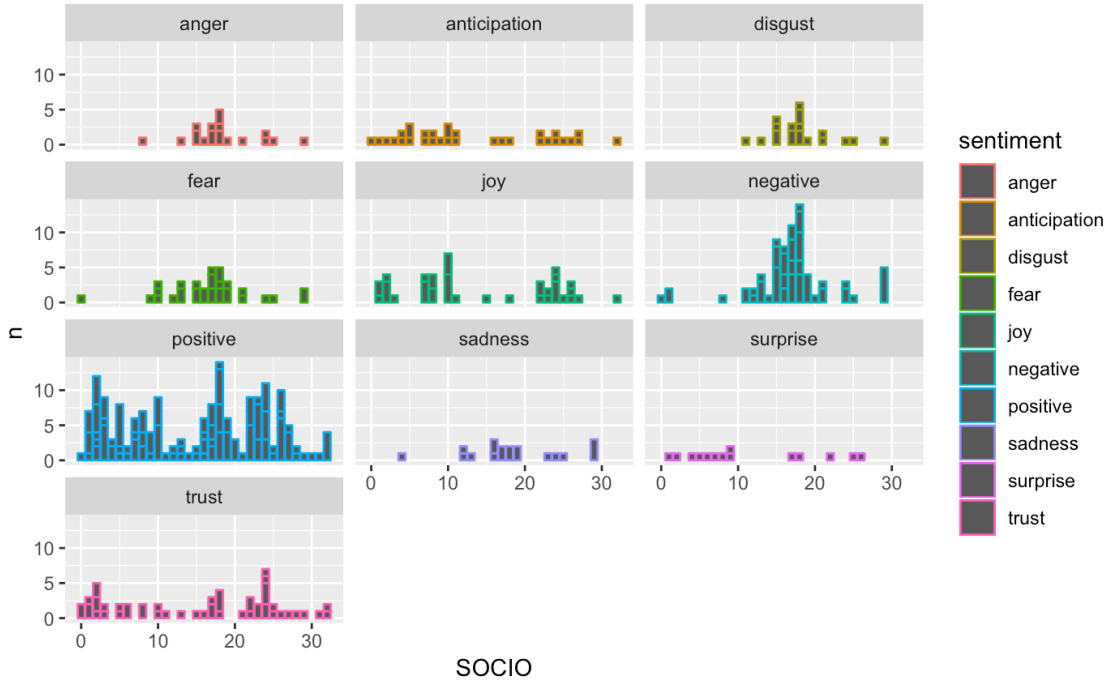

Figure 4.5: Overall satisfaction analysis of open-ended questions for SOCIO

Within the responses, we firstly ranked all word frequency, then we filtered conjunctions, like “de”, “a”, “es”, “y”, “los”, “la”, and classified synonyms to rank the three advantages which were mentioned most.

Table 4.4: Positive word phrases frequency-CREATELY

Positive Aspect	Expression	Count
Quick responsiveness	rápida respuesta/en tiempo real/rápida	15
Easy to use	fácil de usar/fácil de usar/fácil uso /la facilidad de utilizar/fácil de utilizar/facilidad de uso/sencilla	16
Collaborative work	trabajo colaborativo/trabajo en equipo/ trabajo en grupo	10
Friendly Interface	Interfaz amigable/interfaz sencilla/interfaz simple/buena interfaz/interfaz amigable	10
Physical comfort	intuitiva/intuitivo	7
Variety of diagram	variedad de elementos/ variedad de elementos/diferentes opciones de exportación.	4

Table 4.5: Positive word phrases frequency-SOCIO

Positive Aspect	Expression	Count
Collaborative work	trabajo colaborativo/trabajo en equipo/ trabajo en grupo/trabajo en simultáneo	20
Quick responsiveness	rápido/rapidez/rápido/ayuda rápida para manejar diagramas/el tiempo de reacción es muy rápido/en tiempo real/ diagrama rápido	18
Easy to use	fácil/comandos fáciles/facilidad/fácil de utilizar/fácil/fácil de usar	16
Easy control	usable/ tiene documentación para aprender/podemos ver un manual/útil/util	8
Physical comfort	Intuitiva/cambio con vista inmediata/ agradable a la vista	8
Pleasure	entretenido/divertido/interesante	7
Creately	innovadora/novedoso/diseño	7
User-friendly	amigable	4
Learnability	ayuda a no perderse/ la ayuda del chatbot	4
Interactively	interactiva	3
Trustworthy	confiable	3
Communicate easily	lenguaje natural/ lenguaje “normal”/ lenguaje universal	3

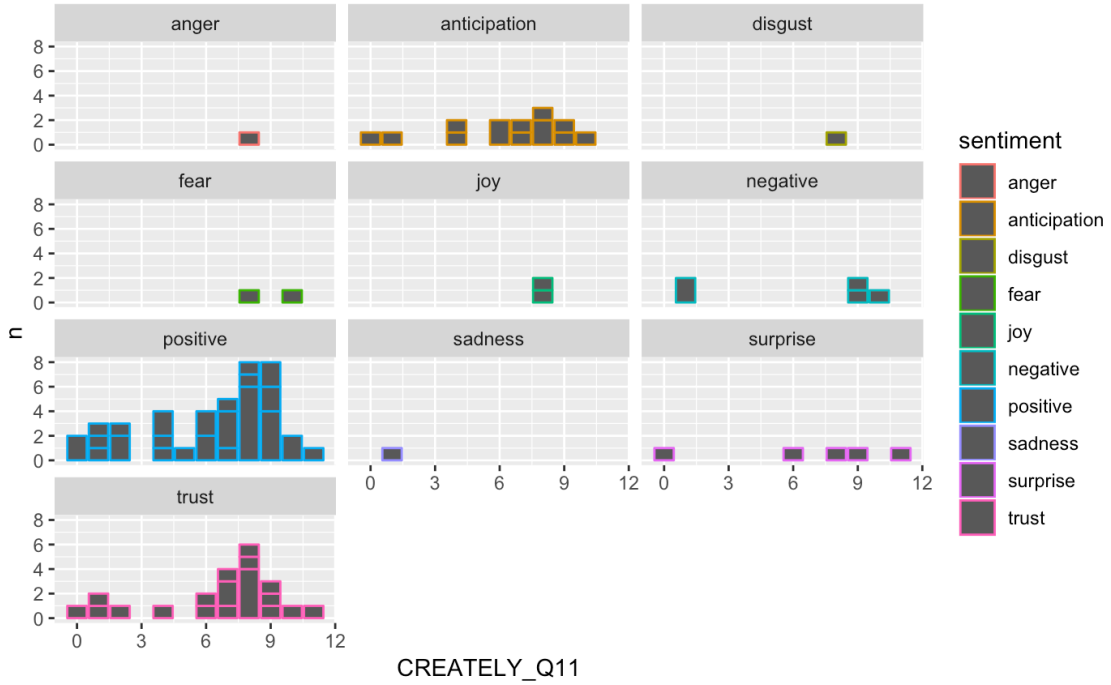

Figure 4.6: Positive aspect of satisfaction analysis for CREATELY

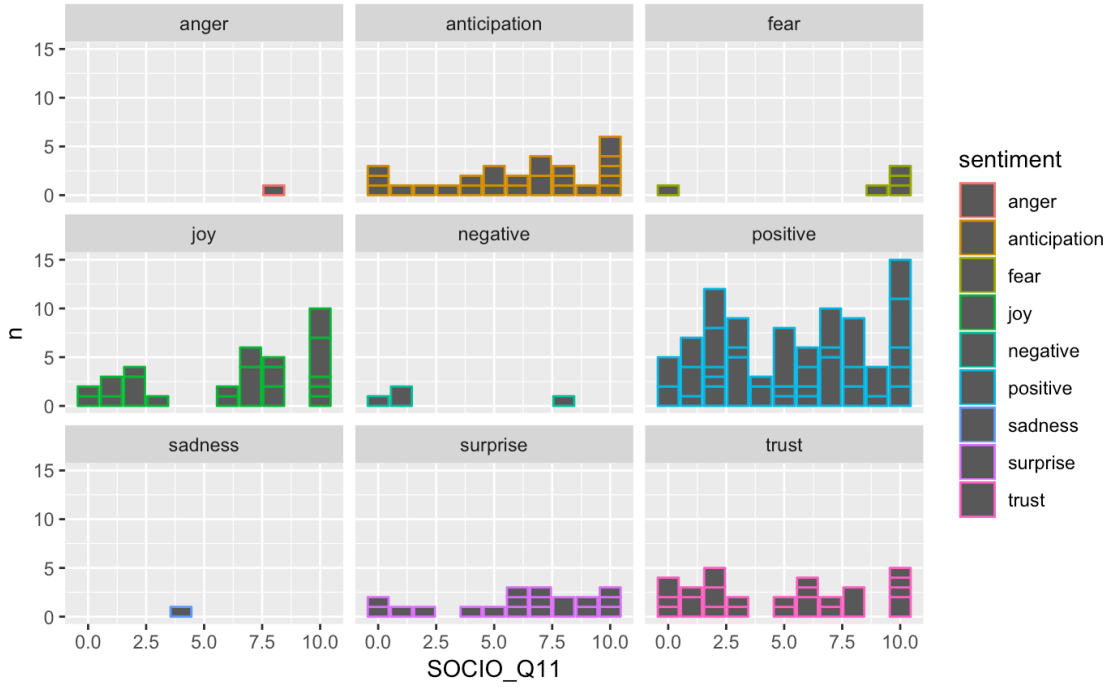

Figure 4.7: Positive aspect of satisfaction analysis for SOCIO

Q: Please indicate three negative aspects of the tool

From Figure 4.8 and Figure 4.9, both SOCIO and CREATELY obtained negative comment, but SOCIO gained more. According to the word frequency, the negative word "Falta" was mentioned 26 times. By reviewing the answer, most of participants who mentioned this word expressed that they are lacking more guide and examples towards the tool, though the developer of chatbot SOCIO had given a guide including example

commands and situations. Meanwhile, emotions of anger, disgust and fear are correspondingly rise.

According to the negative comments, the major problem of CREATELY is that it is hard to realize the real time collaboration. When user loading the application, it occured various of difficult. Some participants were not satisfied with the interface since it is too simple. Besides, its functions are not comprehensive enough. The list of negative comments for CREATELY is shown in Table 4.6. As for SOCIO (see Table 4.7), the limited language of the chatbot SOCIO is its greatest problem. Though the developer of chatbot SOCIO had given a guide including example commands and situations, some participants claimed that the chatbot's user guide needs to be improved and add more examples. Commands of the chatbot SOCIO are not easy enough to learn and the chatbot is lacking some commands. Part of the participants expressed the chatbot SOCIO requires prior related knowledge.

Table 4.6: Negative word phrases frequency-CREATELY

Negative Aspect	Expression	Count
Real time collaboration hard	-Que no se puede editar el mismo elemento dos personas al mismo tiempo como en chart -Al momento de editar una clase si un miembro esta editando el otro no puede trabar con las demás opciones es decir la clase se desactiva o algo parecido. -Lag en la actualización de cambios en los diagramas -Un poco de confusion al momento de agregar colaboradores -Es muy incomo al trabajar en grupo -Necesita un chat incorporado	23
Hard to start	-Incomodidades de inicio de sesión -Necesita adobe flash y la UX se vuelve molesta -Se demora en abrir la página -Se cuelga mucho	23
Physical discomfort	-Interfaz de usuario aburrida -No es agradable para la vista -Colores simples en la interfaz	5
Lack of function	-Los tipos de atributo no existen -Falta varias líneas o relaciones para unir los diagramas -No tiene tanta variedad como otras herramientas	5
Not intuitive	-No es muy intuitiva con el usuario -No es tan intuitiva para quien esta empezando en el proceso de creación de diagramas de clases	4
Lack of auto-save function	-Falta de autoguardado -Log al momento de guardar cambios	4
It's not free	-No es gratuita -Algunas funcionalidades están limitadas cuando no se ha pagado	3

Table 4.7: Negative word phrases frequency-SOCIO

Negative Aspect	Expression	Count
Language limited	-El idioma inglés que se usa para el procesamiento de lenguaje natural presenta ciertas limitantes -Si la persona no maneja un nivel de ingles podría enfrentar complicaciones	12
Guide limited	-La documentación oficial no mostró todas las variantes -Falta más información para la realización de los diagramas -Falta de información de relación	8
Commands limited	-Dificultoso hasta aprender los comandos -Falta de comandos -Toca aprender bien los comandos -No hay comandos para eliminar las relaciones -La forma de hacer relaciones no es tan simple y hay que pensar mucho en como decirlo. -Es confuso el reemplazo de los sustantivos plurales por singulares	7
Must learn it before use	- Al ser la primera vez en usarlo, es dificil adaptarse a los comandos - Es necesario el aprenderse los comandos - Se tiene que tener conocimiento previo de los comandos	7
Cost lots of time	-Se debe eliminar capa por capa cuando existe un error -Tiempo de creación demoroso -Dificultad de borrar	4
Can't correct the error	-No existe detección o corrección de errores para cumplir con la sintaxis de los comandos - Cuando te pierdes literal te pierdes	4
Bugs	-Las excepciones mandan los errores 400 y 500 -Existe un error -Bugs	3

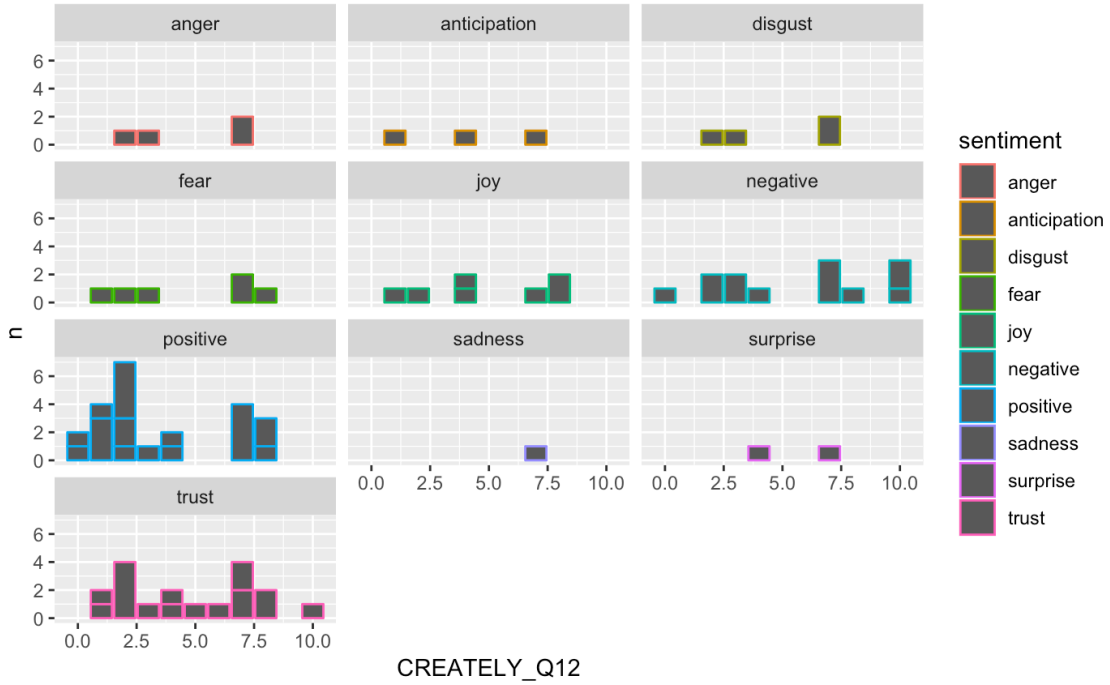

Figure 4.8: Negative aspect of satisfaction analysis for CREATELY

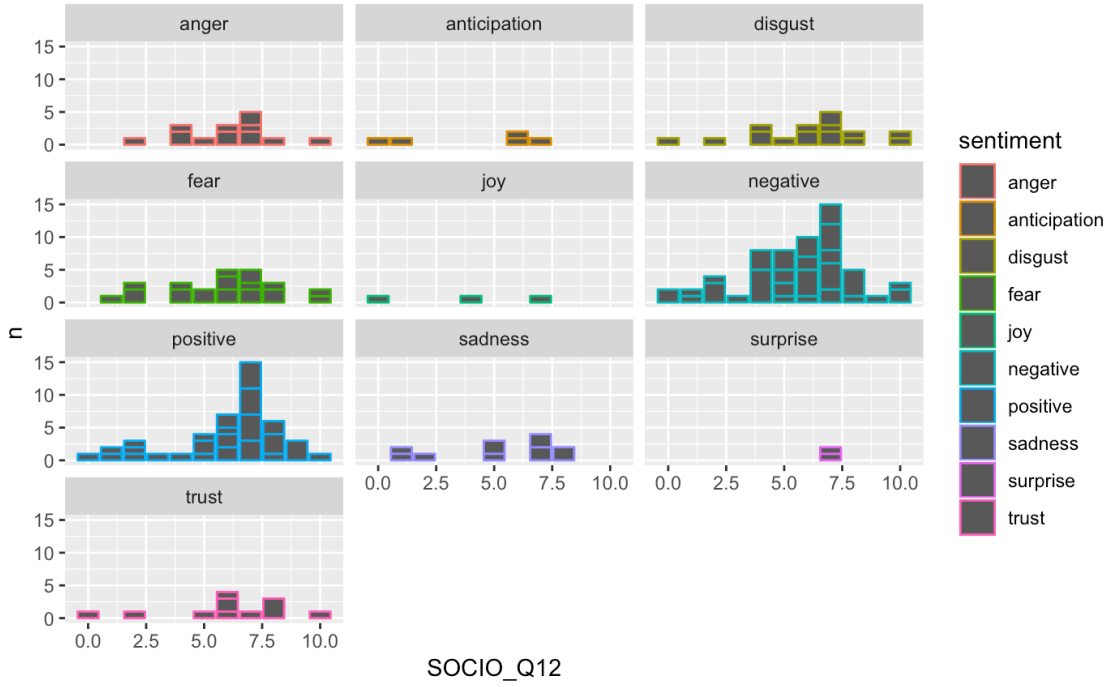

Figure 4.9: Negative aspect of satisfaction analysis for SOCIO

Q: Do you have any suggestions for improvement?

Regarding to Figure 4.10 and Figure 4.11, participants showed relatively positive emotions towards both tools, especially in aspect of anticipation. Besides, they expressed more trust and less sadness for SOCIO than CREATELY.

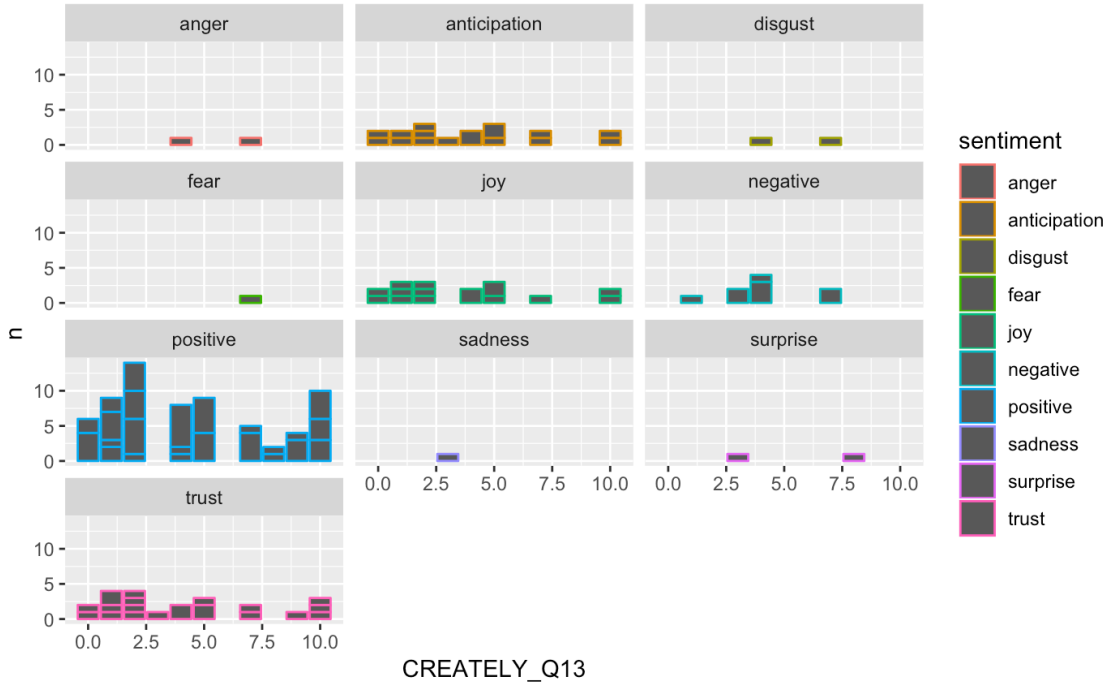

Figure 4.10: Suggestion analysis for CREATELY

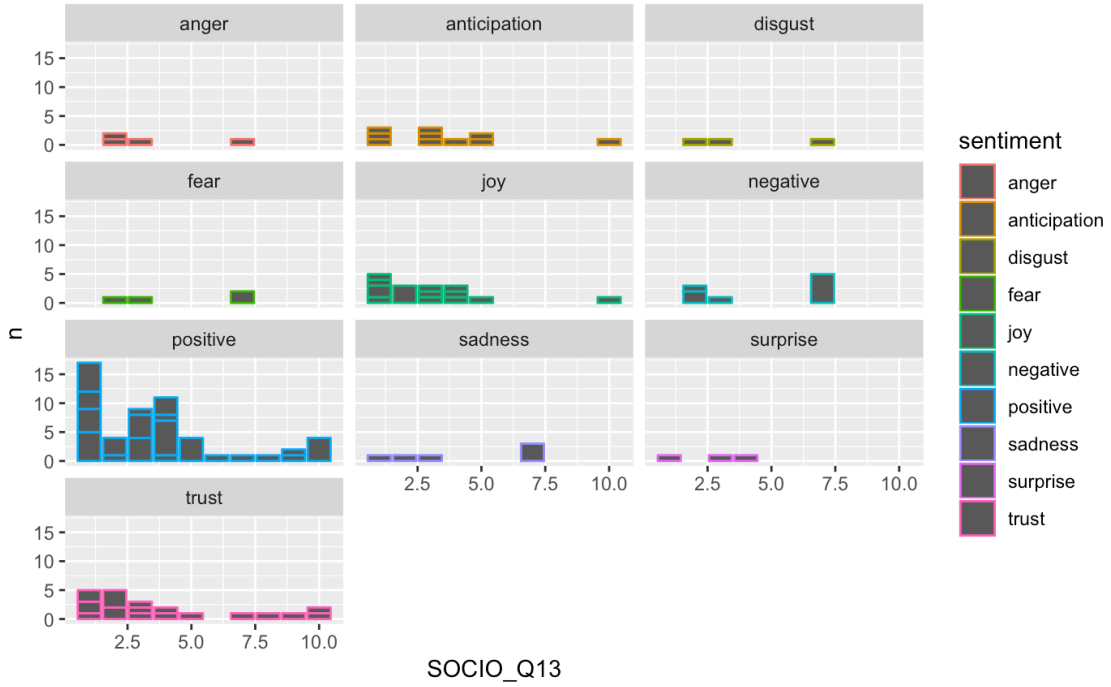

Figure 4.11: Suggestion analysis for SOCIO

Q: Which tool do you prefer?

We count both individual and group preferences. Regarding the individuals, 34 of the participants prefer SOCIO, while 20 of them expressed they prefer CREATELY. Within the groups, 12 groups were more willing to choose SOCIO, the rest of 6 groups favored CREATELY. The ratio of the two tools under these two standards is similar, and in any case, SOCIO has gained more preferences. Individual and group preference for both SOCIO and CREATELY is shown in Figure 4.12 and Figure 4.13.

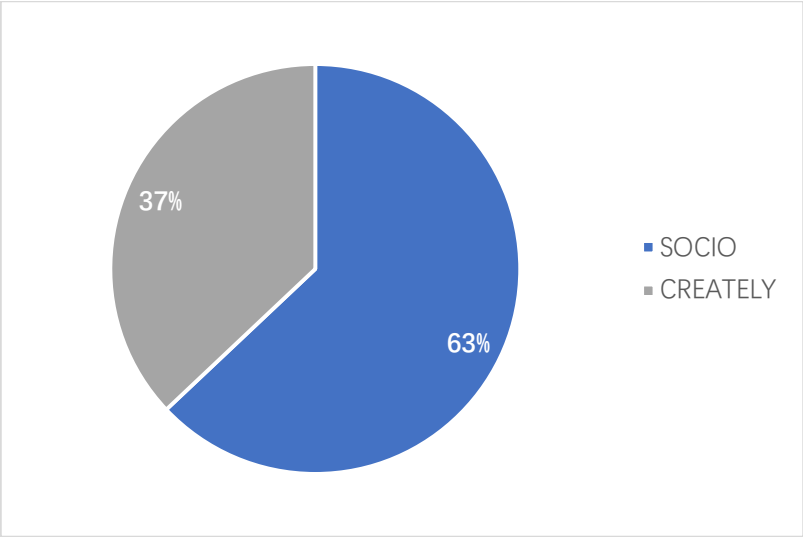

Figure 4.12: Individual preference between SOCIO and CREATELY

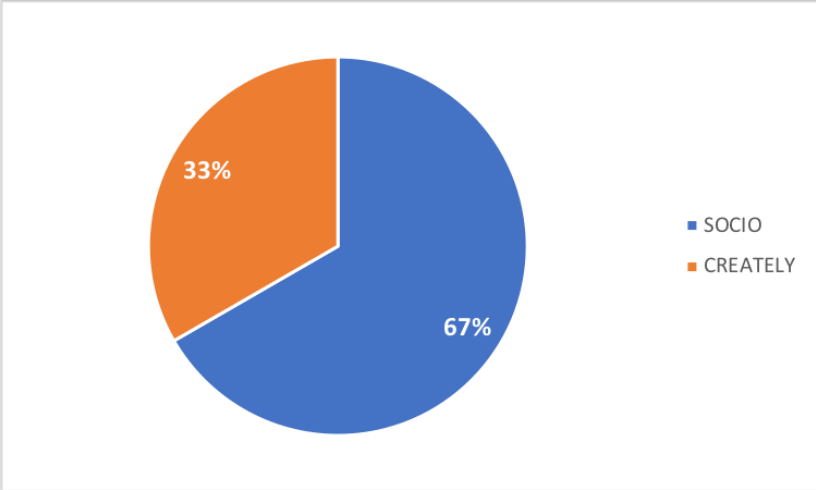

Figure 4.13: Group preference between SOCIO and CREATELY

As for CREATELY, it was suggested to improve real time collaboration. Since some of participants think it is too simple, they suggested improving its interface. In addition, some consider it is should improve its ease-to-use. Table 4.8 details the overall suggestion for CREATELY. Table 4.9 illustrates the overall suggestion for SOCIO. A number of participants advise that it to support more language and add more content into manual.

Table 4.8: Suggestion for CREATELY

Suggestion	Expression	Count
Improve real time collaboration	-Permite editar el mismo elemento varías personas a la vez -Se debe mejorar el trabajo colaborativo	6
Improve interface	-Poner más herramientas al costado de la pantalla -Podrían mejorar la interfaz, el tipo de letra es muy simple -Que la interfaz sea más amigable se ve muy seria	5
Make it easier to use	-Mejorar la manera de editar los diagramas -Al momento de eliminar un diagrama para el resto del grupo les queda un imagen que puede molestar la visualización del diagrama	5
Improve function	-Mejorar el guardado de cambios -Hacer banners informativos con tips de aplicación -Implementar una forma más rápida de hacer las relaciones o asociaciones, -Limitar a valores ya establecido la edición de tipo de campos	4
Improve speed to access	-Dejar de depender de adobe flash, es obsoleto -La rapidez del sitio	4
Improve manual	-Mayor información previa	3
Correct bugs	-Corregir los bugs al trabajar en colaborativo	2
Multilanguage support	-Implementación de otros idiomas	1
Reduce the cost	-Mejorar las funcionalidades free o quizá crear un "plan" para estudiantes con algún precio reducido.	1

Table 4.9: Suggestion for SOCIO

Suggestion	Expression	Count
Multilanguage support	-Tenerla en varios idiomas -Tienen que mejorar el NLP para distintos idiomas	14
Improve manual	-Material de apoyo con ejemplos complejos para poder realizar los diagramas -Añadir más funcionalidades y una guía de inicio rápida -La documentación debe ser más detallada	6
Improve command	-Agregar más comandos -Mejorar reconocimiento de comandos y funciones -Mejorar la manera de enviar los comandos	4
Prediction	-Sugerencia en caso de un comando mal escrito -Una posible sugerencia o predicción de comandos que están incorrectos	3
Improve interaction	Que posea algunos comandos para interactuar más amigables, es decir un saludo cosas así	3
Multiplatform support	Lo hicieran para WhatsApp a lo que estamos más acostumbrados	1
Voice Control	Que reconozca comandos de voz	1
Improve interface	Mejorar interfaz	1

4.4.2. Questions of the SUS

We calculated the SUS score of each of the participant's preference to both tools so that we could quantify this data and make side-by-side comparisons. We adopted Brook's [30] equations to derive the numerical value of each user's individual chatbot preference score. The corresponding equations are shown below:

For questions 1, 3, 5, 7, 9:

$$\text{Sum1} = \text{score value} - 1 \quad (1)$$

For questions 2, 4, 6, 8, 10:

$$\text{Sum2} = 5 - \text{score value} \quad (2)$$

$$\text{SUS score} = 2.5 * (\text{sum1} + \text{sum2}) \quad (3)$$

Based on the values derived from this equation, we compared these two tools in matters of satisfaction. This calculation provided us with a system to quantify satisfaction as if it were an entity capable of being measured. Figure 4.14 shows the box-plot corresponding to the SUS scores of the teams per treatment. As we can see in Figure 4.14, satisfaction scores are higher for SOCIO than CREATELY.

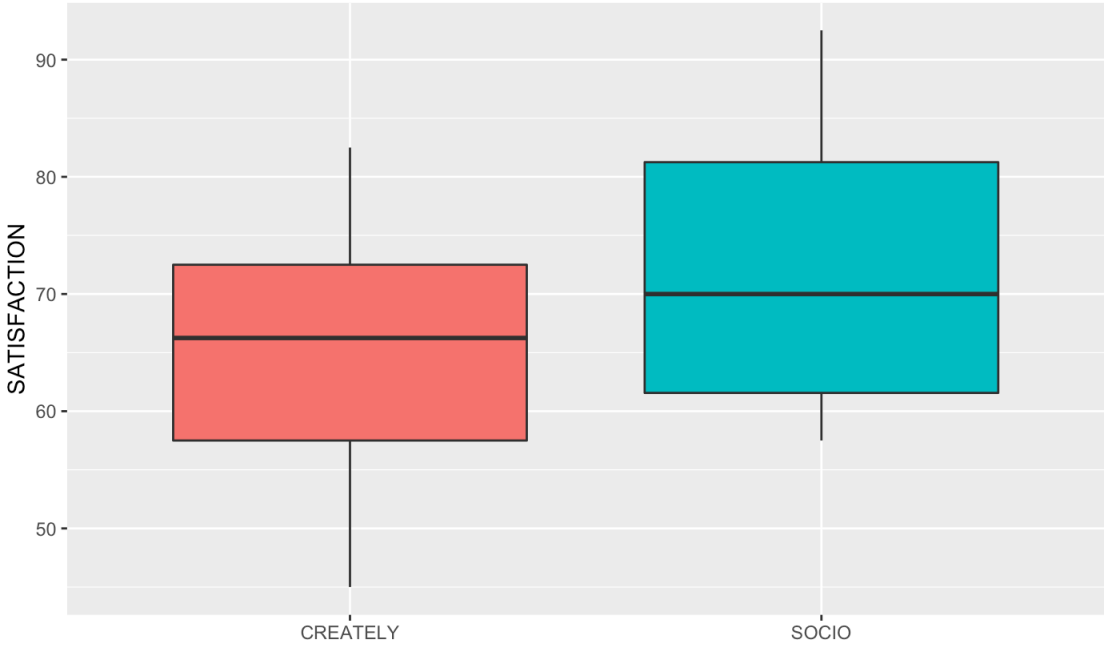

Figure 4.14: Satisfaction scores for CREATELY and SOCIO

Table 4.10 shows the results of the linear mixed model we fitted to analyze the data. As we can see in Table 4.10, neither the sequence nor the order has a statistically significant effect on satisfaction. However, the treatment is quasi-significant. Finally, d=0.58, SE(d)=0.35, suggesting that a medium effect size -according to rules of thumb [4]- materialized for the treatments in terms of satisfaction. This leads us to the conclusion that **SOCIO seems to satisfy users to a greater extent than CREATELY**.

Table 4.10: Linear Mixed Model for satisfaction

	Estimate	**Std. Error**	***p*-value**
(Intercept)	64.51	3.88	0
Seq	1.69	3.97	0.69
Treatment	6.60	3.79	**0.10**
Order	-1.18	3.79	0.75

4.5. Quality Analysis

4.5.1. Precision

Figure 4.15 shows the box-plot corresponding to the precision scores of the teams per treatment. As we can see in Figure 4.15, precision scores are obviously higher for SOCIO than CREATELY. Table 4.11 shows the results of the linear mixed model we fitted to analyze the data. As we can see in Table 4.11, neither the sequence nor the order has a statistically significant effect on precision. However, the treatment is statistically significant at the 0.0002 level. Finally, d=-1.41, SE(d)=0.42, suggesting that a very large effect size -according to rules of thumb [4]- materialized for the treatments in terms of satisfaction. In sum, **SOCIO exceeds CREATELY in terms of precision.**

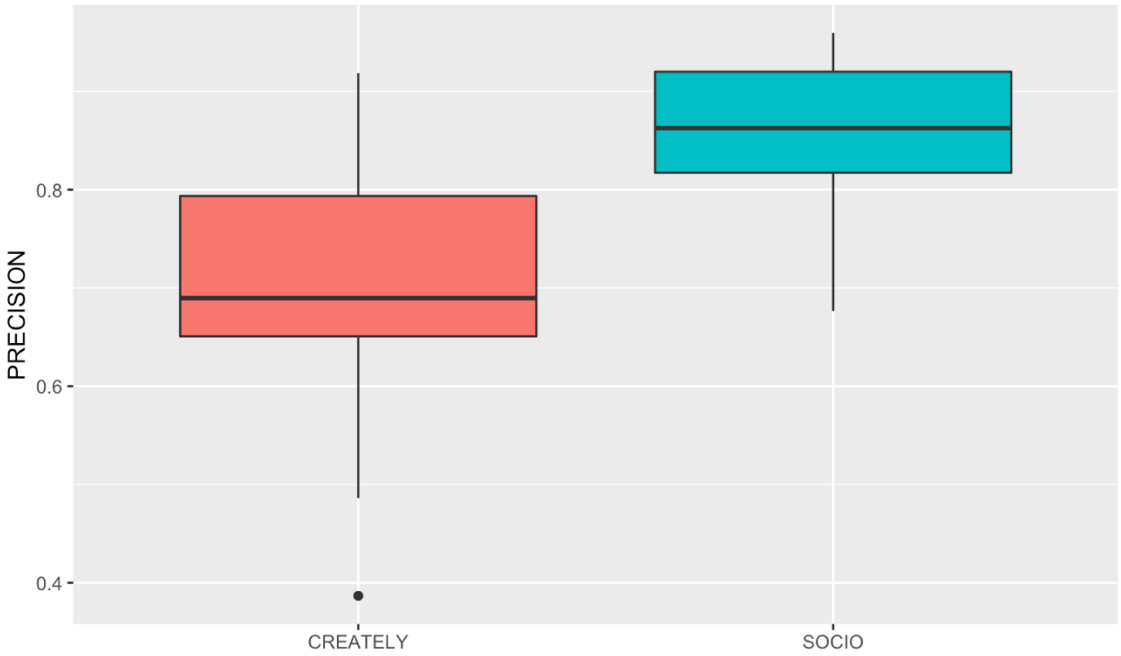

Figure 4.15: Precision scores for CREATELY and SOCIO

Table 4.11: Linear Mixed Model for precision

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.74	0.038	0
Seq	-0.05	0.042	0.25
Treatment	0.16	0.035	**0.0002**
Order	-0.04	0.035	0.24

4.5.2. Recall

Figure 4.16 shows the box-plot corresponding to the recall scores of the teams per treatment. As we can see in Figure 4.16, recall scores are slightly higher for SOCIO than CREATELY. Table 4.12 shows the results of the linear mixed model we fitted to analyze the data.

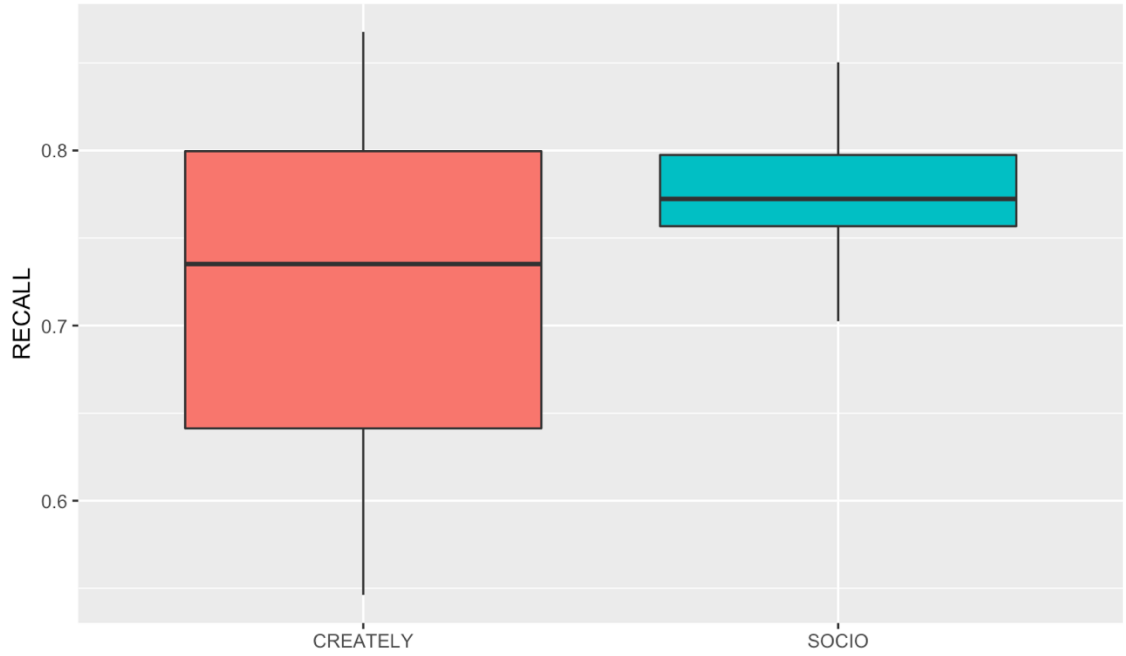

Figure 4.16: Recall scores for CREATELY and SOCIO

Table 4.12: Linear Mixed Model for recall

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.72	0.025	0
Seq	-0.02	0.025	0.25
Treatment	0.057	0.0249	**0.0002**
Order	0.025	0.0249	0.24

As we can see in Table 4.12, only the treatment has a statistically significant effect on recall. It can be stated **SOCIO outperformed CREATELY in terms of recall.** Especially, d=-0.75, SE(d)=0.37, suggesting that a relatively large effect size -according to rules of thumb [4]- materialized for the treatments in terms of accuracy.

4.5.3. Accuracy

Figure 4.17 shows the box-plot corresponding to the accuracy scores of the teams per treatment. As we can see in Figure 4.17, accuracy scores seem higher for SOCIO than for CREATELY. Table 4.13 shows the results of the linear mixed model we fitted to analyze the data.

Figure 4.17: Accuracy scores for CREATELY and SOCIO

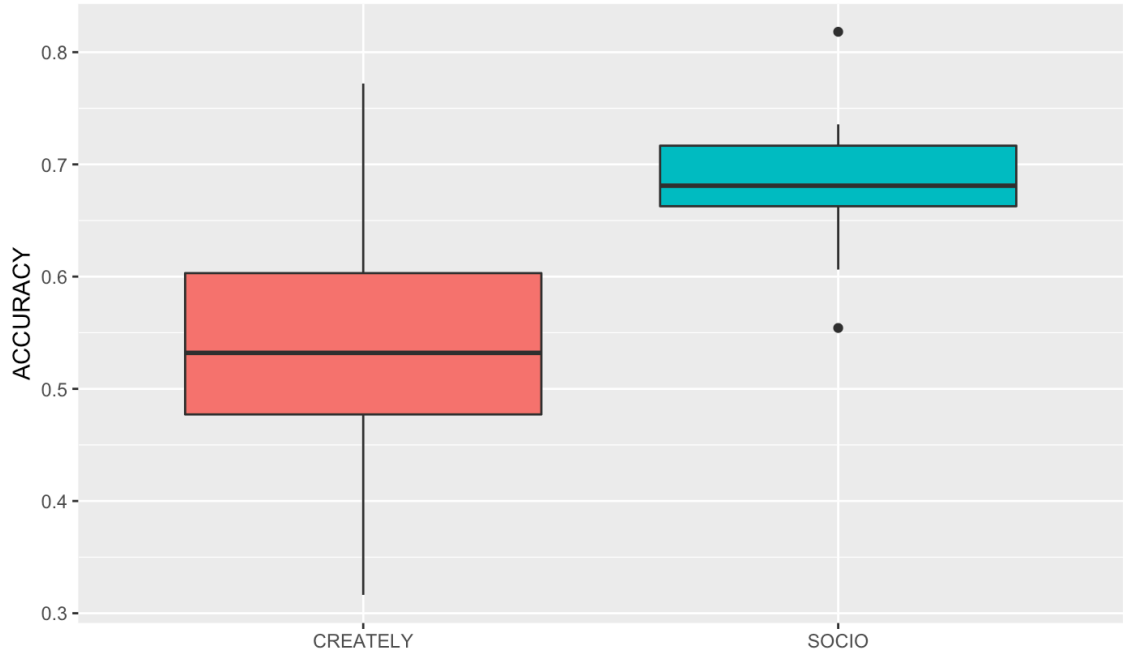

Table 4.13: Linear Mixed Model for accuracy

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.57	0.03	0
Seq	-0.04	0.03	0.22
Treatment	0.14	0.03	**<0.001**
Order	-0.015	0.03	0.64

As we can see in Table 4.13, only the treatment has a statistically significant effect on accuracy. Notably, d=-1.51, SE(d)=0.48, suggesting that a very large effect size -according to rules of thumb [4]- materialized for the treatments in terms of accuracy. **In sum, SOCIO outperformed CREATELY in terms of accuracy.**

4.5.4. Error

Figure 4.18 shows the box-plot corresponding to the error scores of the teams per treatment. As we can see in Figure 4.18, error scores seem apparently lower for SOCIO than for CREATELY. Table 4.14 shows the results of the linear mixed model we fitted to analyze the data.

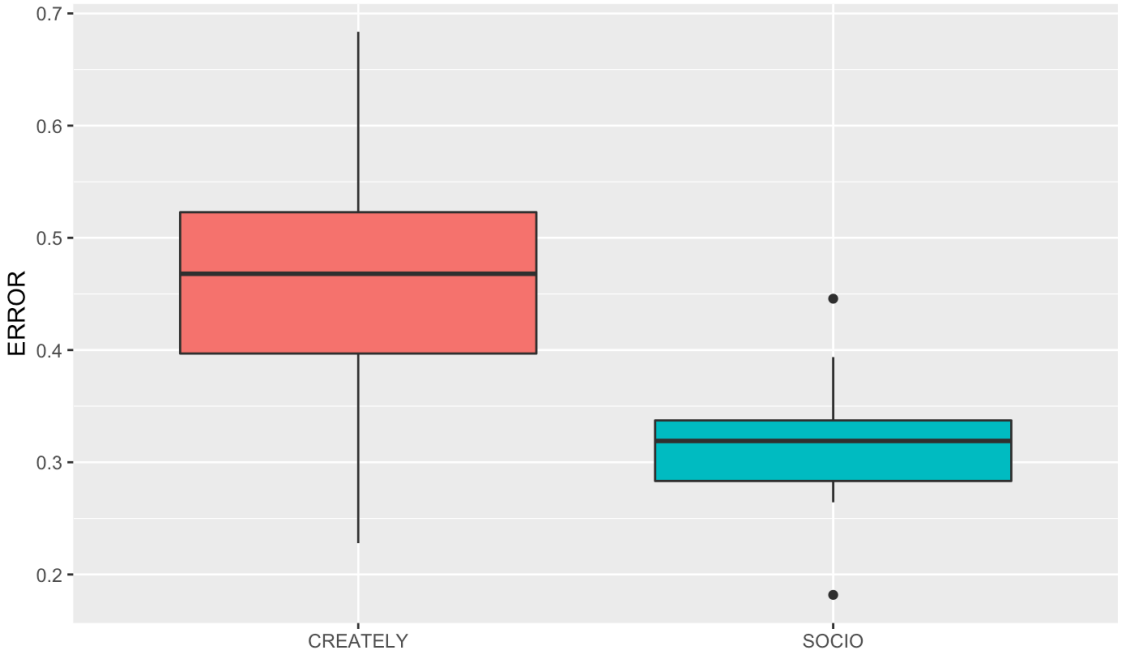

Figure 4.18: Error scores for CREATELY and SOCIO

Table 4.14: Linear Mixed Model for error

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.43	0.031	0
Seq	0.04	0.031	0.22
Treatment	-0.14	0.031	**0.0004**
Order	0.015	0.031	0.64

As we can see in Table 4.14, only the treatment has a statistically significant effect on error. In particular, d=1.50, SE(d)=0.48, suggesting that a very large effect size -according to rules of thumb [4]- materialized for the treatments in terms of error. **In sum, SOCIO occurred less error than CREATELY.**

4.5.5. Success

Figure 4.19 shows the box-plot corresponding to the perceived success scores of the teams per treatment. As we can see in Figure 4.19, perceived success scores seem manifestly higher for SOCIO than for CREATELY.

As we can see in Table 4.15, only the treatment has a statistically significant effect on perceived success. **This demonstrated SOCIO was perceived as more successful than CREATELY.** In particular, d=-1.07, SE(d)=0.41, suggesting that a very large effect size -according to rules of thumb [4]- materialized for the treatments in terms of error. Table 4.15 shows the results of the linear mixed model we fitted to analyze the data.

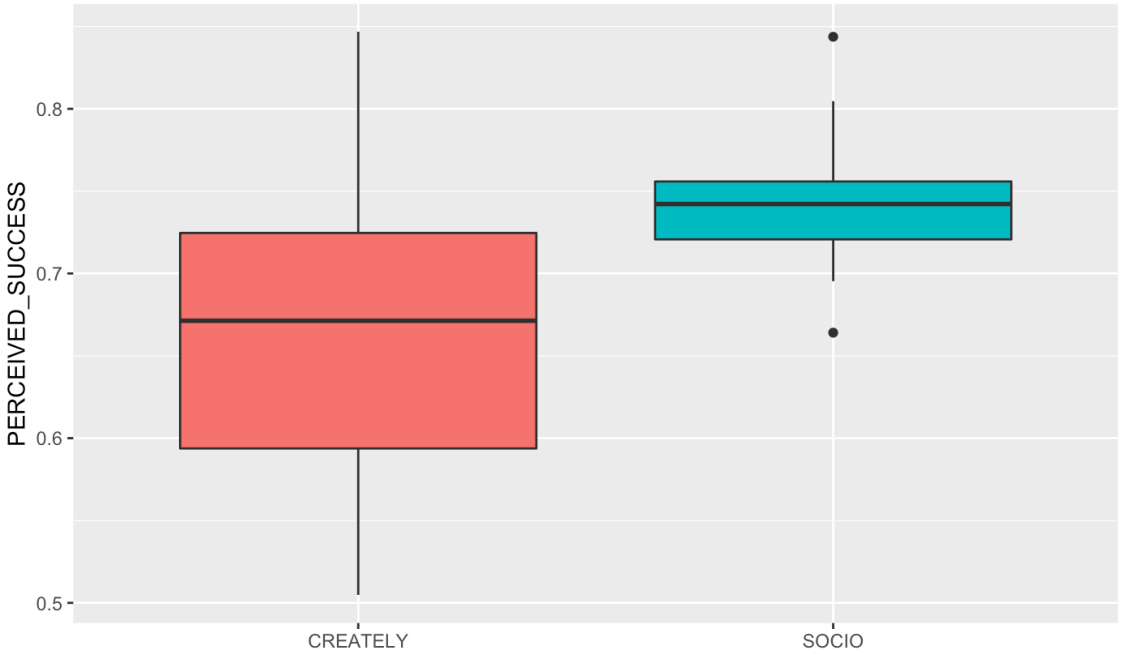

Figure 4.19: Perceived success of CREATELY and SOCIO

Table 4.15: Linear Mixed Model for perceived success

	Estimate	**Std. Error**	***p*-value**
(Intercept)	0.67	0.024	0
Seq	-0.0036990	0.024	0.88
Treatment	0.076	0.02398751	**0.006**
Order	-0.01	0.0239875	0.68

Obviously, it can be concluded from above results analysis of precision, recall, accuracy, error and success, **SOCIO owns higher quality than CREATELY.**

CHAPTER 5

SOCIO ANALYSIS

The number of different interactions of the teams with the chatbot SOCIO while executing the tasks has been analyzed. In particular, a series of independent t-tests have been carried out [12] -one per dependent variable (number of all messages, number of error messages and number of useful messages sent to SOCIO, number of descriptive messages, number of commands and number of actions triggered)- to compare the mean number of the different interactions generated during two tasks (order 1 and 2, respectively). The results of the t-test are complemented with Cohen's d (d, hereinafter) following Borenstein et al.'s formulae [4]. This chapter describes the process of analyzing the experiment data about the chatbot SOCIO. In terms of efficiency, we analyze chatbot SOCIO in aspects of interactivity and fluency.

5.1. Fluency of SOCIO

Fluency of SOCIO is measured by the number of all messages and the number of error messages sent to SOCIO. We define "All messages" as all kinds of messages sent to SOCIO, it includes error messages, valid messages, etc. It should be pointed that command and message sometimes are counted as two messages. For example, the messages "/talk add house", I count it as one message, and the messages "/talk" + "add house", I count them as two messages.

Error messages are those message SOCIO doesn't understand and messages whose intention was to be sent to the chatbot but the participants failed to write correctly, for example, "add house" is an error message, the user should send it like "/talk add house" or "/talk" + "add house", though we understand they wanted to send this command to SOCIO.

5.1.1. Number of All Messages Sent to Chatbot SOCIO

Figure 5.1 shows the box-plot for the number of messages sent to chatbot SOCIO in Task 1 and 2. As we can see in Figure 5.1, the numbers of messages to chatbot SOCIO are similar in Task 1 and Task 2.

As we can see in Table 5.1 the mean number of messages sent to the bot in Task 1 is slightly greater than in Task 2. However, this difference is not statistically significant (p-value=0.87), and a wide confidence interval materialized (95% CI = [-14.96,17.40]). This suggests that **the higher number of messages sent to chatbot SOCIO in Task 1 could be due to random chance alone**. Besides, a small d=0.08 SE(d)=0.22 materialized in the experiment.

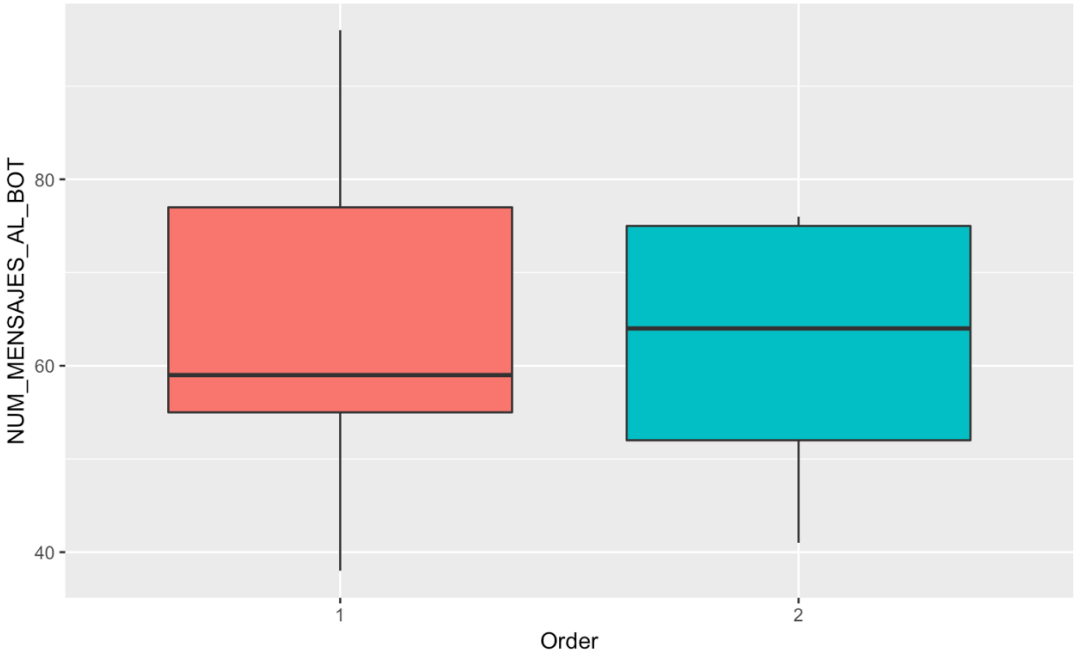

Figure 5.1: Number of all messages to chatbot SOCIO

Table 5.1 shows the results of the t-test comparing the mean number of messages to chatbot SOCIO in Task 1 and 2.

Table 5.1: Mean number of messages to chatbot SOCIO in Task 1 and 2

Task 1	Task 2	95% CI	*p*-value
62.89	61.67	[-14.96,17.40]	0.87

5.1.2. Number of Error Messages Sent to Chatbot SOCIO

Figure 5.2 shows the box-plot for the number of error messages to chatbot SOCIO in Task 1 and 2. As we can see in Figure 5.2, the numbers of error messages to chatbot SOCIO are similar in Task 1 and Task 2.

Table 5.2 shows the results of the t-test comparing the mean number of error messages to chatbot SOCIO in Task 1 and 2. As we can see in Table 5.2 the mean number of error messages is slightly greater in Task 2 than in Task 1. However, again this difference is not statistically significant and a wide 95% CI materialized (95% CI= [-6.78, 6.12]). Finally, a small d=-0.052 SE(d)=0.22 materialized in the number of error messages to chatbot SOCIO. This suggest that since **there is no obvious difference between Task 1 and Task 2, the number of error messages prompted in both tasks is similar.**

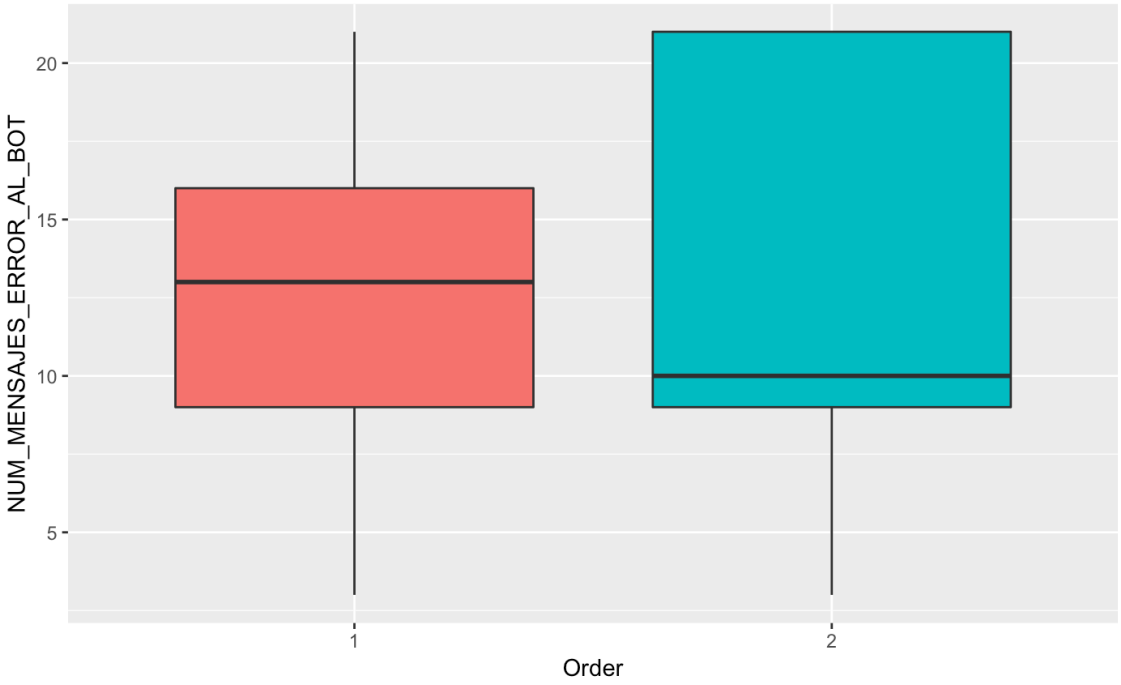

Figure 5.2: Number of error messages to chatbot SOCIO

Table 5.2: Mean number of error messages to chatbot SOCIO in Task 1 and 2

Task 1	Task 2	95% CI	*p*-value
12.67	13.00	[-6.78, 6.12]	0.91

It can therefore be concluded that the fluency of SOCIO is similar in both tasks.

5.2. Interactivity of SOCIO

Interactivity of SOCIO is measured by the number of useful messages, the number of descriptive messages and the number of commands sent to SOCIO and the number of actions triggered by chatbot SOCIO when teams perform Task 1 or Task 2.

The **useful message** is any valid message which can cause actions to SOCIO, it can be descriptive messages or commands. If messages (sent to SOCIO) start with word “add”, “create”, “make”, “remove”, “erase”, “delete” or their synonyms or write like “Attribute + to be + type”, they are defined as **commands.** The rest of useful messages is **descriptive messages**. For example, “/talk There are two doors in a room”.

Action triggered by SOCIO means the change made by the message to SOCIO. It can be counted automatically by the chatbot.

5.2.1. Number of Useful Messages Sent to Chatbot SOCIO

Figure 5.3 shows the box-plot for the number of messages sent to chatbot SOCIO that supposed a contribution to obtain the class diagram. Note that the number of useful messages in Task 2 is greater than in Task 1.

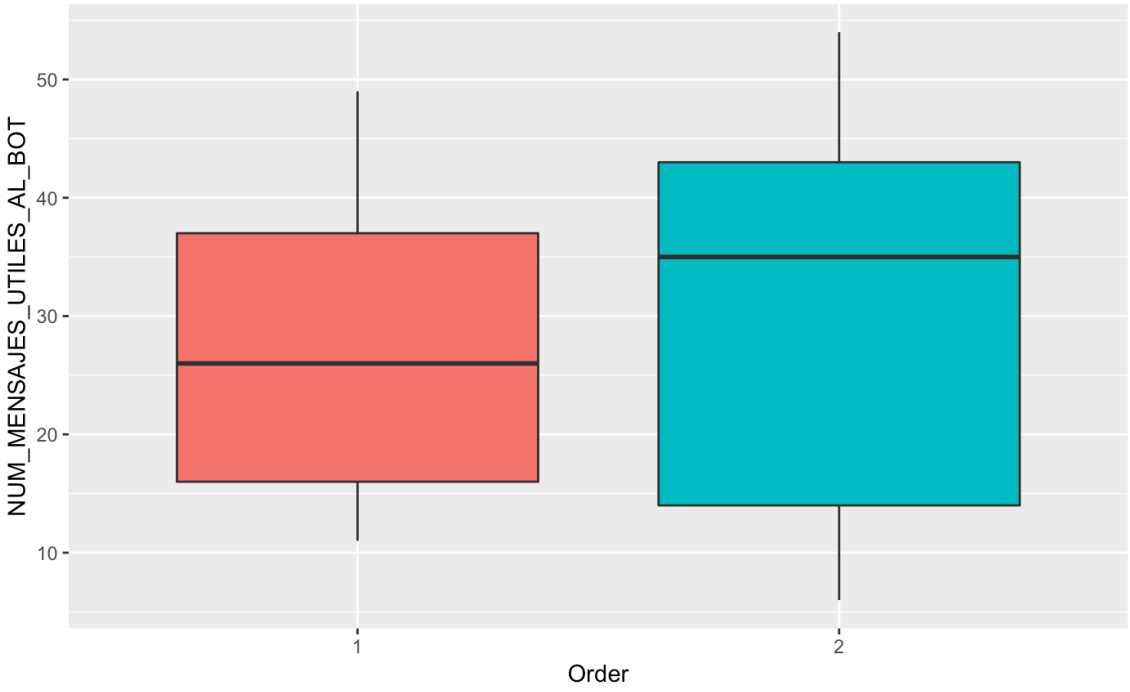

Figure 5.3: Number of useful messages sent to chatbot SOCIO

Table 5.3 shows the result of the t-test comparing the mean number of error messages to chatbot SOCIO in Task 1 and Task 2.

Table 5.3: Mean number of useful messages sent to chatbot SOCIO in Task 1 and 2

Task 1	Task 2	95% CI	*p*-value
28.22	30.22	[-17.78, 13.78]	0.79

As shown in Table 5.3, the mean number of messages sent to SOCIO in Task 2 is greater than in Task 1. However, this difference is not statistically significant (p-value = 0.79), and a broad confidence interval materialized (95% CI = [-17.78, 13.78]). This suggests that **the number of messages directed to the chatbot SOCIO is larger in the second task, may be due to an isolated random cause.** Finally, an average effect size, d = -0.12 and SE (d) = 0.22, is materialized in the number of messages.

5.2.2. Number of Descriptive Messages Sent to Chatbot SOCIO

Figure 5.4 shows the box-plot for the number of descriptive messages sent to chatbot SOCIO that supposed a contribution to the class diagram obtained. It is observed that the number of descriptive messages is similar in both tasks, being slightly higher in Task 1 and more dispersed in Task 2.

Table 5.4 presents the results of the t-test, comparing the average of descriptive messages sent to chatbot SOCIO in Task 1 and 2. As shown in Table 5.4, the mean number of messages sent to chatbot SOCIO in Task 1 is slightly greater than in Task 2. However, this difference is very small (p-value = 0.96), and shows wide range of confidence (95% CI = [-5.12, 5.34]). Finally, a small effect size, d =0.02 and SE (d) = 0.22, is materialized in the number of descriptive messages. **This suggests that since the results are not significant different, the number of descriptive messages generated in both tasks is similar.**

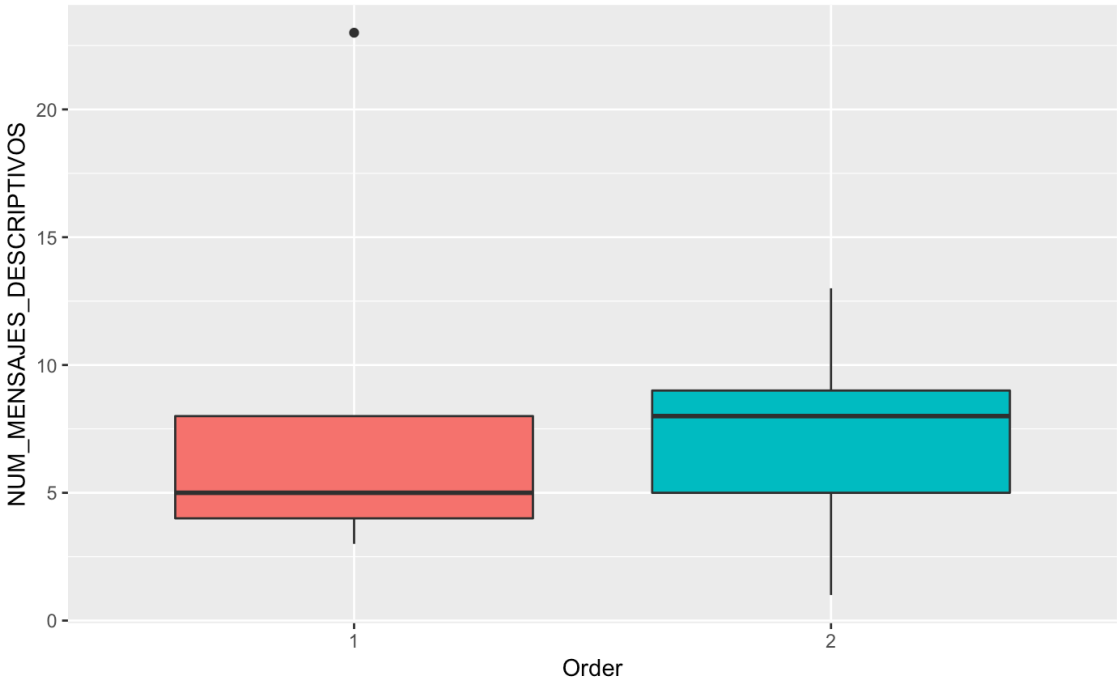

Figure 5.4: Number of descriptive messages sent to chatbot SOCIO

Table 5.4: Mean number of descriptive messages sent to SOCIO in Task 1 and 2

Task 1	Task 2	95% CI	*p*-value
7.22	7.11	[-5.12, 5.34]	0.96

5.2.3. Number of Commands Sent to Chatbot SOCIO

Figure 5.5 shows the box-plot for the number of commands sent to chatbot SOCIO used to construct the class diagram. It is noted that the number of commands sent to Task 2 is slightly greater than to Task 1.

Table 5.5 presents the results of the t-test, comparing the average of numbers of commands sent to chatbot SOCIO in Task 1 and 2. As shown in Table 5.5, the mean number of commands messages sent to chatbot SOCIO in Task 2 is slightly greater than in Task 1. However, this difference is not statistically significant (p-value = 0.74), and a wide range of confidence (95% CI = [-15.91, 11.69]). **This indicates that the number of commands messages generated in both tasks is similar.** Finally, an average effect size, d = -0.15 and SE (d) = 0.22, is materialized in the number of commands.

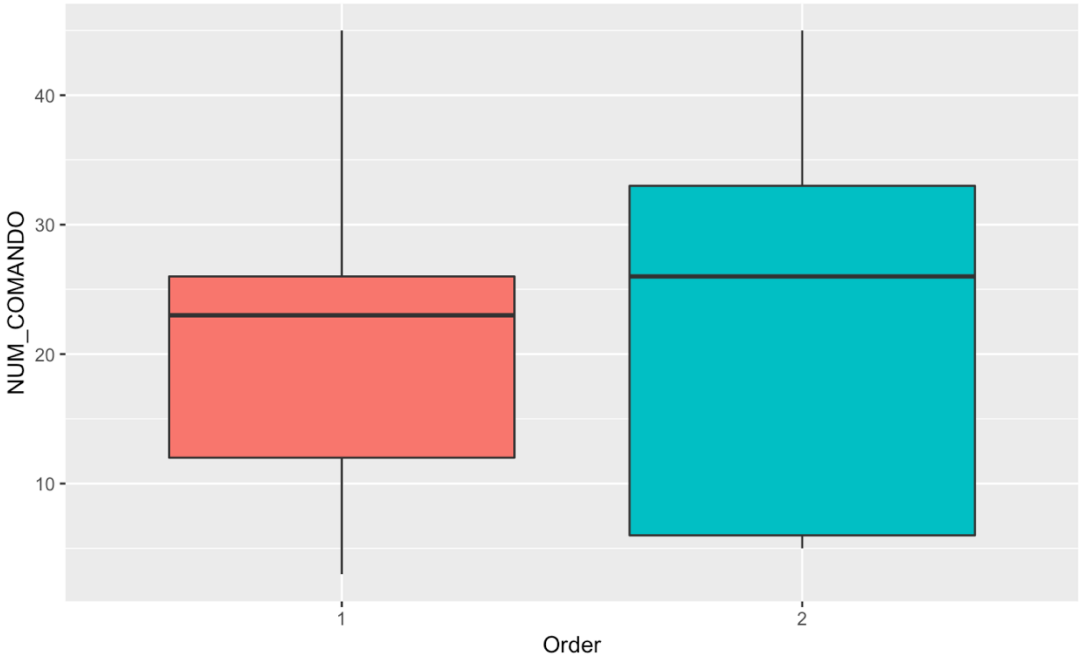

Figure 5.5: Number of commands sent to chatbot SOCIO

Table 5.5: Mean of commands messages sent to chatbot SOCIO in Task 1 and 2

Task 1	Task 2	95% CI	*p*-value
21.00	23.11	[-15.91, 11.69]	0.74

5.2.4. Number of Actions Triggered by Chatbot SOCIO

Figure 5.6 shows the box-plot for the number of actions triggered by chatbot SOCIO during the realization of the class diagram. It is observed that the number of actions triggered in Task 2 is greater than in Task 1.

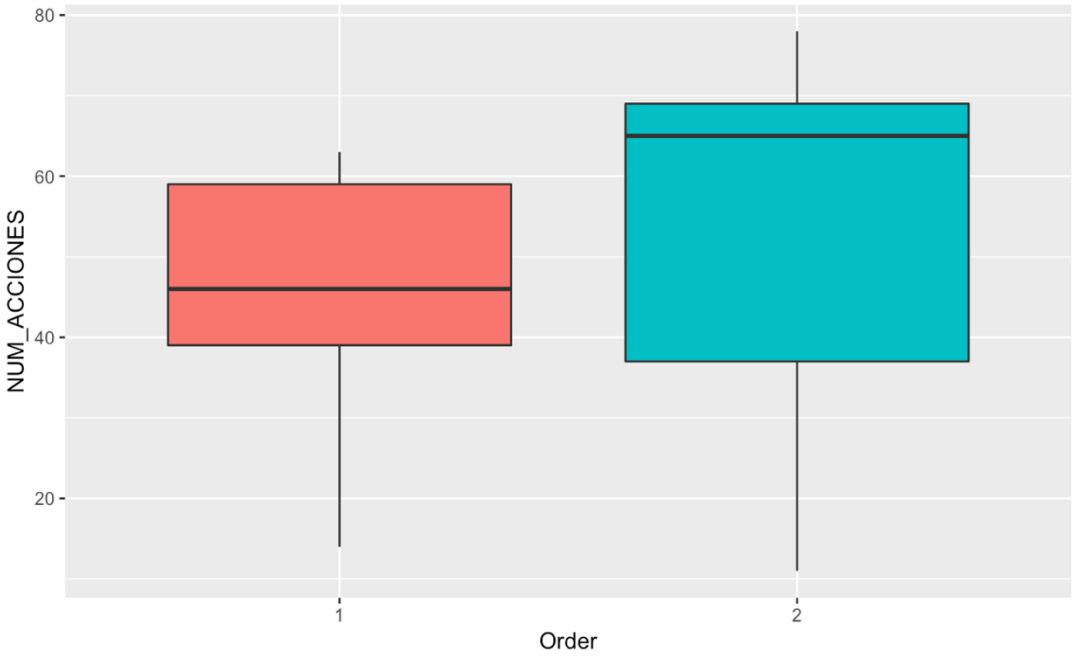

Figure 5.6: Number of actions triggered by chatbot SOCIO

Table 5.6 presents the results of the t-test, comparing the mean number of actions triggered by chatbot SOCIO in Task 1 and 2.

Table 5.6: Mean number of actions triggered by chatbot SOCIO in Task 1 and 2

Task 1	**Task 2**	**95% CI**	***p*-value**
45.56	50.56	[-25.67, 15.67]	0.61

As shown in Table 5.6, the mean number of actions sent to SOCIO in Task 2 is greater than in Task 1. However, this difference is not statistically significant (p-value = 0.61), and a confidence interval of mean amplitude (95% CI = [- 25.67, 15.67]). Finally, an effect size is relatively large, d = -0.25 and SE (d) = 0.22, it materializes in the number of triggered actions. This suggests that despite the fact that the results are not meaningful, **the number of actions triggered in Task 2 is greater than those generated in Task 1.**

In sum, **the interactivity of SOCIO is similar in each task which is proved by all the variables above.**

CHAPTER 6

CONCLUSION

This chapter will conclude the content of all previous chapters in order to recap what has been achieved. It will give a global view of the completed work as well as provide pointers for future work.

6.1. Conclusion

Starting from a subjective definition of the chatbot and usability in the HCI area and our hypothesis for our research, an SMS of chatbot usability were laid down on paper in Chapter 2. Then in Chapter 3, an experimental setting was developed for comparing two tools for creating the class diagram with the chatbot SOCIO and CREATELY. Moreover, the chatbot usability evaluation of both tools was decomposed into several aspects. Of those aspects, some of them correspond to criteria verified in SMS. These four aspects efficiency, effectiveness, satisfaction and quality were analyzed extensively in Chapter 4. The analysis for chatbot SOCIO from interactivity and fluency as shown in Chapter 5. Finally, the final summary and possible improvements were shown in Chapter 6.

RQ1: What is the state of the art of usability in the development of chatbots?

We retrieved 19 primary studies dealing with integration from four different perspectives: usability techniques, usability characteristics, research methods and types of chatbots. The usability techniques are applied to evaluate the usability of the developed chatbot, but not in the analysis and design activities of the chatbot. The procedure that is more frequently followed to evaluate the usability of chatbot is to select a group of subjects to use the chatbot freely or perform certain tasks and then measure satisfaction with a SUS survey.

RQ2: How to evaluate the usability of chatbots using HCI principles?

The evaluation of the usability of chatbots must be done considering the context of use, i.e. the environment where the chatbot will be used, and with representative subjects to whom the chatbot is directed. The most commonly used research methods are surveys, experiments and usability tests. The experimentation and replication of experiments is key within HCI. Achieving successful replications in a discipline, allows its results to be added to those of previous replications, making the body of knowledge grow. However, there is an absence of controlled experiments and replicas measuring usability in chatbots.

There are many ways for practitioners to apply the usability material in this section. The chatbot implementation team can use usability characteristics (Tables 2.6, 2.7 and 2.8) as checklists to help them solve critical problems. Comparing the test results of the same system at different times can check whether the usability characteristics is improved or

decreased. The real-life application of a chatbot will save time to companies, leading to financial gain because of the tasks it is able to take on. As the intelligence and technology of chatbots evolve, they will take on more responsibilities. The chatbot industry is very much interested in the adoption of usability techniques in its development process. On this ground, there is a need for usability-aware design guidelines. Therefore, it is necessary to continue the conclusion of my proposal by experimenting more chatbot usability. This essential work is done in the investigation as followed in Chapter 3, 4 and 5.

RQ3: Does the use of the chatbot SOCIO has a positive effect on the efficiency, effectiveness and satisfaction of the participant when making a class diagram, as well as its quality?

We evaluated usability of chatbot SOCIO from these four aspects: efficiency, effectiveness, satisfaction and quality. For efficiency, the **speed** of SOCIO is exceeded CREATELY since it cost less time when realize each class diagram and it shows **high fluency** and it has an interaction-cost advantage over CREATELY. For effectiveness, SOCIO **outperformed** CREATELY in terms of **completeness**. For satisfaction, SOCIO **satisfies** users to a greater extent than CREATELY with respect to empirical results of SUS score. More users expressed that they **prefer** SOCIO than CREATELY. For quality, SOCIO owns higher inherent quality than CREATELY which is confirmed by analyzing precision, recall, accuracy, error and success. Our experiment results fail to reject the null hypotheses only for SOCIO.

On the one hand, there is no significant difference in the **number of all messages or number of error messages** sent to chatbot SOCIO by the teams when performing Task 1 or Task 2, we can conclude that **fluency of SOCIO in each task is similar.** On the other hand, there is no significant difference in the **number of actions** triggered by chatbot SOCIO when teams perform Task 1 or Task 2 and there is no significant difference in the **number of useful messages OR number of descriptive messages OR number of commands** sent to chatbot SOCIO by the teams when performing Task 1 or Task 2. This shall be without prejudice that **interactivity of SOCIO in each task is similar** too.

6.2. Discussion and Future Work

The analysis in Chapter 2 reveals that the incorporation of usability techniques in the chatbot development process in a formalized manner is strongly represented in the primary studies. We found three papers reviewing the chatbot literature: one discussing the conversational interfaces, patterns, and paradigms [22], one investigating design techniques for conversational agents [40], and a systematic review of conversational agents in healthcare [23]. None of them does an SMS in chatbots usability, which proves our work is original. After that, corresponding with the precious SMS, we conducted an experiment using a within-subject cross-over design to evaluate usability of chatbot SOCIO. We successfully proved that it performed well in terms of efficiency, efficiency, satisfaction and quality compared with web-based application CREATELY.

Judging by the increase in publications since 2015, the integration of usability of chatbots is of notable interest. However, there is no agreement on what would be a formalized and more systematic integration yet. Therefore, it is an open research problem that requires more research effort. Even though the literature retrieved by the SMS provides a picture

of chatbot usability, no paper provides generally applicable guidelines for chatbots usability.

On one hand, the validity of the SMS reported in this work is threatened by including only papers written in English. On the other hand, the authors of an SMS may make errors of judgement when analyzing the relevant publications. This is a horizontal rather than a vertical analysis, on which ground relevant publications may have been overlooked. Additionally, although the terms used in the search string were the most commonly accepted by other authors, other terms used describing relevant publications may have been overlooked. Finally, the publications were evaluated and classified based on the judgment and experience of the authors, and other researchers may have evaluated the publications differently.

The chatbot SOCIO used for our study had limited capabilities, which participants pointed out. Participants highlighted improving the understanding capability (NLP) of SOCIO by supporting more different languages (namely, Spanish), exploring more social media platforms to launch this chatbot, adding more example and details into manual, and auto-correcting spelling mistakes. Nevertheless, the goal of this study was to evaluate the usability of SOCIO, a chatbot without such advanced features sufficed. In fact, the satisfaction with our participants is relatively high.

Results from our experiment provide evidence that users who haven't used these two tools before were able to become productive in 30 minutes using the chatbot to realize the task of creating the class diagram. The experiment was unable to confirm whether the more sophisticated class diagram creation capability is helpful as this capability was used only to a very limited extent by participants, most likely due to time pressures of the task. Besides, we plan on conducting a second round of evaluations engaging more users to interact with chatbot SOCIO, especially we will aim at more English native speakers.

REFERENCES

[1] Acquire™. “Top 11 Chatbots Trends to Keep an Eye on in 2019 | Acquire”. [online] Available at: https://acquire.io/blog/chatbots-trends. 2017. [Accessed 13/06/2019].

[2] A. Alghamdi, M. Owda and K. Crockett. “Natural language interface to relational database (NLI-RDB) through object relational mapping (ORM)”. Advances in Intelligent Systems and Computing, vol. 513. Springer, Cham. 2017.

[3] L. Au. “Chat conversation methods traversing a provisional scaffold of meanings: U.S. Patent Application 11/806,261”. 2007.

[4] M. Borenstein, L.V. Hedges, J.P. Higgins and H.R. Rothstein. “Introduction to meta-analysis”. John Wiley & Sons. 2011.

[5] Botnation.ai. “Chatbot A/B Testing Optimizations.”[online] Available at: http:// help.botanation.ai/articles/1471932-chatbot-a-b-testing-optimizations.2019. [Accessed 12/06/2019]

[6] Chatbots Journal. “Chatbots and their Impact on the User Experience (UX)”. [online] Available at: https://chatbotsjournal.com/chatbots-and-their-impact- on-the-user-experience-ux-f200d906c1a0. 2019 [Accessed 13/06/2019].

[7] Chatbots Magazine. “Usability Heuristics for Bots”. [online] Available at: https://chatbotsmagazine.com/usability-heuristics-for-bots-7075132d2c92. 2019. [Accessed 13/06/2019].

[8] M.L. Chen and H.C. Wang. “How personal experience and technical knowledge affect using conversational agents”. In Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, p. 53. 2018.

[9] A. Cheng, V. Raghavaraju, J. Kanugo, Y.P. Handrianto and Y. Shang. “Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes”. In Proceedings of the 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC’18), pp. 1-5. 2018.

[10] N.C. Chi, O. Sparks, S.Y. Lin, A. Lazar, H.J. Thompson and G. Demiris. “Pilot testing a digital pet avatar for older adults”. Geriatric Nursing, 38(6), pp. 542-547. 2017.

[11] D. Elmasri and A. Maeder. “A conversational agent for an online mental health intervention”. In Proceedings of the International Conference on Brain and Health Informatics, pp. 243-251. 2016.

[12] A. Field, J. Miles and Z. Field. “Discovering Statistics Using R”. SAGE, 2012.

[13] A. Følstad and P.B. Brandtzæg. “Chatbots and the new world of HCI”. Interactions, 24(4). pp. 38-42. 2017.

[14] Forbes. “How Chatbots Will Transform Customer Experience: An Infographic”. [online] Available at: https://www.forbes.com/sites/blakemorgan/2017/03/21/

how-chatbots-will-transform-customer-experience-an-infographic/#3be671597fb4. 2017 [Accessed 13/06/2019].

[15] J.P. Higgins and S. Green. “Cochrane handbook for systematic reviews of interventions”. Chichester, UK: John Wiley & Sons. 2008.

[16] K. Hornbæk. “Current practice in measuring usability: Challenges to usability studies and research”. International Journal of Human-Computer Studies, 64(2), pp.79-102. 2006.

[17] ISO 9241-11. “Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)–Part II Guidance on Usability”. 1998.

[18] ISO/IEC 25010. “Systems and Software Engineering–Systems and Software Quality Requirements and Evaluation (SQuaRE)–System and Software Quality Models”. 2010.

[19] M. Jain, R. Kota, P. Kumar and S.N. Patel. “Convey: Exploring the Use of a Context View for Chatbots”. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 468. 2018.

[20] D. Jurafsky and J. H. Martin. “Dialog systems and chatbots”. Speech and Language Processing. pp. 1-26. 2018.

[21] B.A. Kitchenham, D. Budgen and O. Pearl Brereton. “Using Mapping Studies as the Basis for Further Research-A Participant-Observer Case Study”. Information and Software Technology, 53(6), pp. 638–651. 2011.

[22] L.C. Klopfenstein, S. Delpriori, S. Malatini and A. Bogliolo. “The rise of bots: A survey of conversational interfaces, patterns, and paradigms”. In Proceedings of the 2017 Conference on Designing Interactive Systems, pp. 555-565. 2017.

[23] L. Laranjo, A.G. Dunn, H.L. Tong, A.B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A.Y. Lau and E. Coiera. “Conversational Agents in Healthcare: A Systematic Review”. Journal of the American Medical Informatics Association, 25(9), pp.1248-1258. 2018.

[24] J. Lester, K. Branting and B. Mott. “Conversational agents”. The Practical Handbook of Internet Computing, Chapman and Hall/CRC, pp. 220-240. 2004.

[25] LOLA Asistente Virtual de la Universidad de Murcia para Ayudar en el Proceso de Preinscripción y Matrícula. Online: https://amp.laopiniondemurcia.es/comunidad/2018/09/30/lola-convence/958842.html. 2018 [Accessed 14/12/2018].

[26] I. Lopatovska, K. Rink, I. Knight, K. Raines, K. Cosenza, H. Williams, P. Sorsche, D. Hirsch, Q. Li, and A. Martinez. “Talk to me: Exploring user interactions with the Amazon Alexa”. Journal of Librarianship and Information Science, pp. 1-14. 2018.

[27] C. Messina. “2016 Will Be the Year of Conversational Commerce”. Online: https://medium.com/chris-messina/2016-will-be-the-year-of-conversational-commerce-1586e85e3991. 2016 [Accessed 14/12/2018].

[28] J.A. Micoulaud-Franchi, P. Sagaspe, E. De Sevin, S. Bioulac, A. Sauteraud and P. Philip. “Acceptability of embodied conversational agent in a health care context”. In Proceedings of the International Conference on Intelligent Virtual Agents, pp. 416-419. 2016.

[29] S.M. Mohammad and P.D. Turney. “Crowdsourcing a word–emotion association lexicon.” Computational Intelligence, 29(3), pp. 436-465. 2013.

[30] Q.N. Nguyen and A. Sidorova. “Understanding user interactions with a chatbot: A self-determination theory approach”. In Proceedings of the Americas Conference on Information Systems 2018: Digital Disruption. 2018.

[31] A.I. Niculescu, K.H. Yeo, L.F. D'Haro, S. Kim, R. Jiang and R.E. Banchs. “Design and evaluation of a conversational agent for the touristic domain”. In Proceedings of the APSIPA, pp. 1-10. 2014.

[32] J. Nielsen. “Enhancing the explanatory power of usability heuristics”. In ACM CHI'94 Conf. (Boston, MA, April 24-28), pp. 152-158. 1994.

[33] D. Novick and L.M. Rodríguez. “Extending empirical analysis of usability and playability to multimodal computer games”. In Proceedings of the International Conference of Design, User Experience, and Usability, pp. 469-478. 2016.

[34] J. Pereira and O. Díaz. “A quality analysis of Facebook messenger’s most popular Chatbots”. In Proceedings of the ACM SAC Conference (SAC’18), pp. 2144-2150. 2018.

[35] J. Pérez, Y. Sánchez, F.J. Serón, and E. Cerezo. “Interacting with a semantic affective ECA”. In Proceedings of the International Conference on Intelligent Virtual Agents, pp. 374-384. 2017.

[36] K. Petersen, R. Feldt, S. Mujtaba and M. Mattsson. “Systematic mapping studies in software engineering”. In Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, pp. 71-80. 2008.

[37] A. Preece, W. Webberley, D. Braines, E.G. Zaroukian and J.Z. Bakdash. “SHERLOCK: Experimental evaluation of a conversational agent for mobile information tasks”. IEEE Transactions on Human-Machine Systems, 47(6), pp.1017-1028. 2017.

[38] N. M. Radziwill and M. C. Benton. “Evaluating quality of chatbots and intelligent conversational agents”. arXiv preprint arXiv:1704.04579. 2017.

[39] A. M. Rahman, A. Al Mamun and A. Islam. “Programming challenges of chatbot: Current and future prospective”. In Proceedings of the 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 75-78. 2017.

[40] K. Ramesh, S. Ravishankaran, A. Joshi and K. Chandrasekaran. “A survey of design techniques for conversational agents”. In Proceedings of the International Conference on Information, Communication and Computing Technology, pp. 336-350. 2017.

[41] J. Saenz, W. Burgess, E. Gustitis, A. Mena and F. Sasangohar. “The Usability Analysis of Chatbot Technologies for Internal Personnel Communications”. In Proceedings of the 67th Annual Conference and Expo of the Institute of Industrial Engineers (IIE’17). pp. 1357-1362. 2017.

[42] C. Sinoo, S. van der Pal, O.A.B. Henkemans, A. Keizer, B.P. Bierman, R. Looije and M.A. Neerincx. “Friendship with a robot: Children’s perception of similarity between a robot’s physical and virtual embodiment that supports diabetes self-management”. Patient Education and Counseling, 101(7), pp.1248-1255. 2018.

[43] L. Sullivan. “Facebook Chatbots Hit 70% Failure Rate as Consumers Warm Up to the Tech”. Online: https://www.mediapost.com/publications/article/ 295718/2017. [Accessed 14/12/2018].

[44] S. Tegos, S. Demetriadis and T. Tsiatsos. “A configurable conversational agent to trigger students’ productive dialogue: A pilot study in the CALL domain”. International Journal of Artificial Intelligence in Education, 24(1), pp.62-91. 2014.

[45] The Interaction Design Foundation. “Usability Evaluation”. [online] Available at: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/usability-evaluation. 2019. [Accessed 13/06/2019].

[46] The Interaction Design Foundation. “What is Usability?”. [online] Available at: https://www.interaction-design.org/literature/topics/usability. 2019. [Accessed 13/06/2019].

[47] M.L. Tielman, M.A. Neerincx, R. Bidarra, B. Kybartas and W.P. Brinkman. “A therapy system for post-traumatic stress disorder using a virtual agent and virtual storytelling to reconstruct traumatic memories”. Journal of Medical Systems, 41(8), p.125. 2017.

[48] C. Tsiourti, J. Quintas, M. Ben-Moussa, S. Hanke, N.A. Nijdam and D. Konstantas. “The CaMeLi Framework—A multimodal virtual companion for older adults”. In Proceedings of SAI Intelligent Systems Conference, pp. 196-217. 2018.

[49] S. Vegas, C. Apa and N. Juristo. “Crossover designs in software engineering experiments: Benefits and perils”. IEEE Transactions on Software Engineering, 42(2), pp. 120-135. 2016.

[50] A. V. Woudenberg. “A Chatbot Dialogue Manager-Chatbots and Dialogue Systems: A Hybrid Approach”. Master's Thesis. Open University of the Netherlands. Faculty of Management, Science and Technology. 2014.

[51] R. Yaghoubzadeh, K. Pitsch and S. Kopp. “Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users”. In Proceedings of the International Conference on Intelligent Virtual Agents, pp. 28-38. 2015.

APPENDICES

APPENDIX A
PRIMARY STUDY

Table A.1 lists the primary studies located during the mapping study described in this work.

Table A.1: Primary Studies

ID	Source	Title	Author	Year	Type
1	Scopus	Understanding user interactions with a chatbot: A self-determination theory approach	Nguyen, Q.N., Sidorova, A.	2018	Conference Paper
2	Scopus	Talk to me: Exploring user interactions with the Amazon Alexa	Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., Martinez, A.	2018	Article in Press
3	Scopus	How personal experience and technical knowledge affect using conversational agents	Chen, M.-L., Wang, H.-C.	2018	Conference Paper
4	Scopus	Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes	Cheng, A., Raghavaraju, V., Kanugo, J., Handrianto, Y.P., Shang, Y.	2018	Conference Paper
5	Scopus	Convey: Exploring the use of a context view for chatbots	Jain, M., Kota, R., Kumar, P., Patel, S.	2018	Conference Paper
6	Scopus	Friendship with a robot: Children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management	Sinoo, C., van der Pal, S., Blanson Henkemans, O.A., Keizer, A., Bierman, B.P.B., Looije, R., Neerincx, M.A.	2018	Article
7	Scopus	Interacting with a semantic affective ECA	Pérez, J., Sánchez, Y., Serón, F.J., Cerezo, E.	2017	Conference Paper
8	Scopus	Natural language interface to relational database (NLI-RDB) through object relational mapping (ORM)	Alghamdi, A., Owda, M., Crockett, K.	2017	Book Chapter
9	Scopus	A Therapy System for Post-Traumatic Stress Disorder Using a Virtual Agent and Virtual Storytelling to Reconstruct Traumatic Memories	Tielman, M.L., Neerincx, M.A., Bidarra, R., Kybartas, B., Brinkman, W.-P.	2017	Conference Paper
10	Scopus	The CaMeLi framework—A multimodal virtual companion for older adults	Tsiourti, C., Quintas, J., Ben-Moussa, M., Hanke, S., Nijdam, N.A. and Konstantas, D.,	2017	Article
11	Scopus	Sherlock: Experimental Evaluation of a Conversational Agent for Mobile Information Tasks	Preece, A., Webberley, W., Braines, D., Zaroukian, E.G., Bakdash, J.Z.	2017	Article
12	Scopus	Pilot testing a digital pet avatar for older adults	Chi, N.-C., Sparks, O., Lin, S.-Y., Lazar, A., Thompson, H.J., Demiris, G.	2017	Conference Review
13	Scopus	The usability analysis of chatbot technologies for internal personnel communications	Saenz, J., Burgess, W., Gustitis, E., Mena, A., Sasangohar, F.	2016	Conference Paper
14	Scopus	A conversational agent for an online mental health intervention	Elmasri, D., Maeder, A.	2016	Conference Paper
15	Scopus	Acceptability of embodied conversational agent in a health care context	Micoulaud-Franchi, J.-A., Sagaspe, P., De Sevin, E., Bioulac, S., Sauteraud, A., Philip, P.	2016	Conference Paper
16	Scopus	Extending empirical analysis of usability and playability to multimodal computer games	Novick, D., Rodríguez, L.M.	2016	Conference Paper
17	Scopus	Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users	Yaghoubzadeh, R., Pitsch, K., Kopp, S.	2015	Conference Paper
18	Scopus	Design and evaluation of a conversational agent for the touristic domain	Niculescu, A.I., Yeo, K.H., D'Haro, L.F., Kim, S., Jiang, R., Banchs, R.E.	2014	Article
19	Scopus	A configurable conversational agent to trigger students' productive dialogue: A pilot study in the CALL domain	Tegos, S., Demetriadis, S., Tsiatsos, T.	2014	Conference Paper

APPENDIX B
TYPES OF CHATBOTS

Table B.1: Types of Chatbots

Chatbots: Design Techniques and Usability Evaluation Techniques								
Year	Program name	Designer Name	Description	Design Techniques	User Experience	Usability Techniques	Usability Characteristics	Research Methods
2014	SARA	Niculescu, A.I., Yeo, K.H., D' Haro, L.F., Kim, S., Jiang, R., Banchs, R.E.	SARA offers a comfortable solution for those who want to explore the city by themselves and have no tourist guide around	Structured database, Text document, Artificial Intelligence Markup Language (AIML)	According to the feedback of user experience, the system is good, but the venues and direction were not properly identified by the system	Questionnaire	**Effectiveness:** Recall, Accuracy; **Satisfaction:** Complexity Control	Survey, Usability test, Quasi-Experiment
2014	Amazon Alexa	Amazon Company	Echo/Alexa is reportedly used for playing music, answering general questions, setting alarms and timers, or controlling networked devices	Intelligent Personal Assistance (IPA)	High frequencies of satisfaction	Questionnaire	**Effectiveness:** Task completion; **Satisfaction:** Pleasure	Survey
2014	MentorChat	Tegos, S., Demetriadis, S., Tsiatsos, T.	A dialogue-based system that employs a configurable and domain-independent conversational agent for triggering students' productive dialogue.	Computer-supported collaborative learning (CSCL), a cloud-based application	Regarding user acceptance and usability of the system some positive students' opinions were identified	Questionnaire, Interview	**Effectiveness**: Accuracy; **Satisfaction**: Ease-of-use, Pleasure	Survey
2015	An autonomous spoken dialogue assistant	Yaghoubzadeh, R., Pitsch, K., Kopp, S.	A prototype of an autonomous spoken dialogue assistant to support people with age-related or congenital cognitive impairments in the domain of week planning	The flexdiam Dialogue Manager	The evaluation study they have conducted with older adults showed that the general design of the system is suitable to almost match the results obtained with the WOz version of the assistant	Interview	**Effectiveness**: Accuracy	Usability test
2016	Verbot	Saenz, J., Burgess, W., Gustitis, E., Mena, A., Sasangohar, F.	A chatbot for an industry partner to better their internal communication process between field technicians and engineers. Implementation of this technology will automate the technician-to-engineer communication process and thus will result in a much more efficient system	Three chatbot platforms (IBM, Pandora, Self-development Kit)	The results suggest that IBM's Watson represents the technology best aligned with their human factors analysis	Interview, Think-Aloud, Questionnaires (SUS), Cognitive Walkthrough	**Effectiveness**: Task completion; **Satisfaction**: Ease-of-use, Context-dependent question, Learnability	Survey, Usability Test
2016	An Embodied Conversational Agents (ECA) in health care	Micoulaud-Franchi, J.-A., Sagaspe, P., De Sevin, E., Bioulac, S., Sauteraud, A., Philip, P.	An ECA performing a clinical structured interview	ECA	It was showed that patients perceived globally the acceptability of the ECA higher than the tablet	Interview	**Effectiveness**: Task completion	Usability Test
2016	A Conversational Agent for an Online Mental Health Intervention	Elmasri, D., Maeder, A.	Chatbots for a mental health intervention, specifically in alcohol drinking habits assessment	AIML	Obtain positive reception of the intervention by users	Questionnaire, Interview	**Effectiveness:** Experts and Users' assessment, Accuracy; **Satisfaction:** Ease-of-use	Survey
2016	Adriana: a computer game "Survival on Jungle Island"	Novick, D., Rodríguez, L.M.	The players interacted with a life-sized embodied conversational agent (ECA) named Adriana who had been stranded in the island shortly before the player arrived. Adriana would partner with the player to survive on and escape from the deserted island	Embodied Conversational Agent (ECA), Microsoft Kinec	No mention overall user experience, analyze frustration episodes can be viewed as both playability and usability problems	Direct Observation, Interview	**Efficiency:** Task completion time; **Effectiveness**: Accuracy; **Satisfaction**: Context-dependent question, Pleasure	Case Study
2016	The PAL Projec	Sinoo, C., van der Pal, S., Blanson Henkemans, O.A., Keizer, A., Bierman, B.P.B., Looije, R., Neerincx, M.A.	A physical (robot) + a virtual avatar inMyPAL app,support diabetes self-management of children. In MyPAL application, various modules estimating the child's emotional state, monitoring the goal progress and deciding on which action the child should take next. The PAL Projec could play diabetes games and quizzes with the children, teaching them about diabetes and diabetes self-management	Physical (robot): Wizard-of-Oz (WoZ technique)	Users prefer physical robot to virtual avatar	Questionnaire (SUS), Questionnaire (ad-hoc)	**Effectiveness**: Experts and Users' assessment, Accuracy; **Satisfaction**: Ease-of-use, Want to use again, Pleasure, Learnability	Experiment, Survey
2017	E-VOX	Pérez, J., Sánchez, Y., Serón, F.J., Cerezo, E.	E-VOX works as an assistant to provide useful information from Wikipedia, supporting real-feel human-computer interaction	Emotional model; Soar (cognitive-affective architecture); 3D ECA (multimodal interface)	When the ECA has a negative personality, it tends to ignore or blame the user, and gives a stronger emotional reaction than an agent with positive personality that is generally nice and helps them	Questionnaire	**Efficiency**: Communication effort, Task completion time; **Effectiveness**: Recall, Accuracy; **Satisfaction**: Ease-of-use, Want to use again, Pleasure	Experiment

Table B.1: Types of Chatbots (Cont)

Chatbots: Design Techniques and Usability Evaluation Techniques								
Year	Programm Name	Designer Name	Description	Design Techniques	User Experience	Usability Techniques	Usability Characteristics	Research Methods
2017	CaMeLi	Tsiourti, C., Quintas, J., Ben-Moussa, M., Hanke, S., Nijdam, N.A. and Konstantas, D.,	CaMeLi autonomous conversational agent system simulates human-like affective behaviour and acts as a companion for older adults living alone at home	Conversational Agent (CA)	Positive user satisfaction	Questionnaire (SUS), Semi-Structured Interviews	**Satisfaction**: Ease-of-use, Context-dependent question, Before use, Complexity Control	Case Study
2017	SHERLOCK	Preece, A., Webberley, W., Braines, D., Zaroukian, E.G., Bakdash, J.Z.	In SHERLOCK, individuals acted in groups to discover and report information to the machine using natural language (NL), which the machine then processed into controlled natural language (CNL)	Web app in JavaScript to run in Web browsers on a variety of devices and processed in controlled natural language (CNL)	Conversational agent had high usability, while added speech act capabilities would not increase usability	Questionnaire (SUS), Direct Observation	**Efficiency**: Communication effort, Task completion time; Effectiveness: Task completion	Case Study, Quasi-Experiment
2017	A Digital Pet Avatar	Chi, N.-C., Sparks, O., Lin, S.-Y., Lazar, A., Thompson, H.J., Demiris, G.	As population of older adults grows, social support needs also increase. ECAs have the potential to provide such social support for older adults	ECA	The digital pet can provide older adults with companionship and enhance social interaction. However, the agent's conversational ability, technical issues, privacy, and dependence are some issues that need to be addressed.	Questionnaire (SUS), Interviewed with audio recorded	**Efficiency**: Communication effort; **Effectiveness**: Experts and Users' assessment; **Satisfaction**: Before use, During use, Physical discomfort, Want to use again, Pleasure	Survey, Usability Test
2017	A Natural Language Interface to a Relational Database (NLI-RDB)	Alghamdi, A., Owda, M., Crockett, K.	The developed NLI-RDB system allows the user to interact with objects directly in natural language and through navigation, rather than by using SQL statements	Conversational Agent (CA), Information Extraction (IE) and Object Relational Mapping (ORM) framework	Evaluation results showed an excellent user satisfaction in both the quantitative and qualitative evaluations used.	Questionnaire	**Effectiveness**: Recall, Accuracy; **Satisfaction**: Ease-of-use, Physical discomfort	Usability test
2017	3MR_2 system	Tielman, M.L., Neerincx, M.A., Bidarra, R., Kybartas, B., Brinkman, W.-P.	It is desgined for post-traumatic stress disorder (PTSD) patiens. With this system, patients can recollect their memories in a digital diary and recreate them in a 3D WorldBuilder	ECA. The 3MR_2 system has two main exposure environments, a digital diary and a 3D tool, the WorldBuilder.	Initial evaluations revealed that the system was usable by both non-patients and former PTSD patients. These evaluations did reveal small usability concerns, which were then resolved	Think-aloud Questionnaire(SUS)	**Satisfaction**: Ease-of-use	Survey, Experiment, Usability Test
2018	The Healthy Coping in Diabetes Application	Cheng, A., Raghavaraju, V., Kanugo, J., Handrianto, Y.P., Shang, Y.	The central feature of our application focuses on supporting healthy coping in diabetes self-management (DSM). The application also include a monitoring survey that allows users to simply input the information to the bot via voice without having to deal with the organization of that information	Google Home (IPA) + Webhook & Internal Logic + Web Interface	More users are satisfied and have a comfortable user experience	Questionnaire, Interview	Satisfaction: Ease-of-use, Before use, Complexity Control, Physical discomfort	Experiment, Case Study
2018	Hipmunk Chatbot	Nguyen, Q.N., Sidorova, A.	An AI-powered virtual travel-planning assistant that helps search for flight and accommodation options and gives advice and recommendations	Natural Language Processing (NLP)	No mention	Questionnaire	**Efficiency**: Mental effort; **Satisfaction**: During use	Experiment
2018	Apple Siri	Apple Company	Participants were asked to perform a series of trip planning tasks (e.g., "find an inexpensive hotel in Osaka") using only Apple Siri on a mobile phone in 30 minutes	Intelligent Personal Assistance (IPA)	Results showed that prior use experience and technical knowledge impact how people feel about CAs in real use contexts differently	Think-aloud,Interviewed	**Effectiveness**: Experts and Users' assessment; **Satisfaction**: Learnability	Experiment
2018	An e-commerce chatbot added Convey (CONtext View)	Jain, M., Kota, R., Kumar, P., Patel, S.	Convey (CONtext View), a window added to the chatbot interface that displays the (inferred and assumed) context of the con- versation to the user,it also provides intuitive interactions on the context values, enabling users to modify them in a simple and efficient manner	IBM Conversation platform	Participants preferred using chatbot with Convey and found it to be easier to use, less mentally demanding, intuitive	Questionnaire, Interview	**Efficiency**: Mental effort; **Effectiveness**: Task completion; **Satisfaction**: Want to use again, Pleasure	Experiment

APPENDIX C
USABILITY DATA

Table C.1: Usability Data

Usability Data													
				Efficiency							Effectiveness	Satisfaction
				Speed	Fluency		Interactivity					
Group	Team	Task	Tool	Time (min)	#Discussion Messages	#All Messages to Chatbot	#Invalid Messgaes	#Messages Directed to Chatbot	#Descriptive Messages	#Command	#Actions	Completeness	SUS Score
1	1	Task1	SOCIO	30	41	13	2	11	8	3	45	1	85
	2			25	27	30	4	26	3	23	46	1	70
	3			22	2	53	16	37	7	30	61	1	82.5
	4			22	17	28	10	18	5	13	36	0.955	65
	5			26	17	23	8	15	3	12	39	0.97	72.5
	6			27	0	38	4	34	8	26	47	0.985	85
	7			27	0	6	0	6	0	6	11	1	57.5
	8			27	3	7	1	6	1	5	11	1	62.5
	9			29	8	20	6	14	8	6	37	0.985	60
	1	Task2	CREATELY	30	68							1	80
	2			30	11							0.985	57.5
	3			30	7							1	85
	4			30	45							1	47.5
	5			30	7							0.985	52.5
	6			27	9							1	57.5
	7			30	9							1	67.5
	8			25	10							1	65
	9			26	10							0.97	70
2	10	Task1	CREATELY	30	16							0.97	57.5
	11			29	37							0.97	45
	12			30	6							0.97	57.5
	13			30	20							0.955	55
	14			30	8							1	82.5
	15			27	19							1	67.5
	16			29	21							1	85
	17			30	20							0.955	77.5
	18			26	29							1	72.5
	10	Task2	SOCIO	30	7	48	10	38	3	35	65	1	85
	11			30	2	64	21	43	10	33	78	1	77.5
	12			27	1	26	15	11	6	5	21	1	57.5
	13			29	8	52	3	49	4	45	63	1	75
	14			30	0	28	3	25	5	20	37	0.97	67.5
	15			25	12	75	27	48	23	25	59	1	57.5
	16			25	1	75	21	54	9	45	69	1	92.5
	17			26	3	76	30	46	13	33	72	0.97	61.25
	18			30	24	56	21	35	9	26	65	0.985	70

APPENDIX D
QUALITY DATA

Table D.1: Quality Metrics

					Quality Metrics																			
					Number of the clase				Number of the attribute				Number of the relations					Element						
	Tool	Group	Team	Task	Value	Existence	Name	Type	Value	Existence	Type	Value	Existence	Type	cardinality	Value	Value	Value	Succes	Precision	Recall	Accuracy	Error
1	SOCIO	1	1	Task1	TP	6	6	5	5.75	12	12	12	6	6	6	6	23.75	1	0.7421875	0.95959596	0.748031	0.725190844	0.274809
					FP	0	0	0	0	0	0	0	1	1	1	1	1						
					FN	1	0	1	1	6	6	6	1	1	1	1	8						
			2		TP	6	6	5	5.75	14	9	12.75	5	4	5	4.75	23.25		0.7265625	0.93	0.768595	0.726562	0.273438
					FP	0	0	1	0.25	0	5	1.25	0	2	0	0.25	1.75						
					FN	1	1	1	1	4	4	4	2	2	2	2	7						
			3		TP	6	6	6	6	17	13	16	4	3	4	3.75	25.75		0.8046875	0.865546218	0.830645	0.735714	0.264286
					FP	1	1	1	1	2	6	3	0	0	0	0	4						
					FN	1	1	1	1	1	1	1	3	4	3	3.25	5.25						
			4		TP	5	5	4	4.75	18	10	16	3	3	3	3	23.75		0.7421875	0.826086957	0.798319	0.683453237	0.316547
					FP	1	1	1	1	1	9	3	1	1	1	1	5						
					FN	2	2	2	2	0	0	0	4	4	4	4	6						
			5		TP	5	5	4	4.75	15	14	14.75	5	5	5	5	24.5		0.765625	0.816666667	0.777778	0.662162	0.337838
					FP	0	0	1	0.25	5	6	5.25	0	0	0	0	5.5						
					FN	2	2	2	2	3	3	3	2	2	2	2	7						
			6		TP	7	7	6	6.75	16	13	15.25	5	5	5	5	27		0.84375	0.955752212	0.850394	0.818181818	0.181818
					FP	0	0	1	0.25	0	0	0	1	1	1	1	1.25						
					FN	0	0	0	0	2	5	2.75	2	2	2	2	4.75						
			7		TP	5	5	4	4.75	14	10	13	4	4	4	4	21.75		0.6796875	0.945652174	0.707317	0.679688	0.320313
					FP	0	0	1	0.25	0	4	1	0	0	0	0	1.25						
					FN	2	2	2	2	4	4	4	3	3	3	3	9						
			8		TP	6	6	5	5.75	12	9	11.25	7	5	7	6.5	23.5		0.734375	0.870370370	0.770492	0.691176	0.308824
					FP	1	1	2	1.25	0	3	0.75	1	3	1	1.5	3.5						
					FN	1	1	1	1	6	6	6	0	0	0	0	7						
			9		TP	6	6	5	5.75	12	8	11	4	2	4	3	19.75		0.6171875	0.686956522	0.711712	0.537415	0.462585
					FP	2	2	3	2.25	3	2	2.75	3	2	3	4	9						
					FN	1	1	1	1	4	2	2.75	2	2	2	2	8						
1	CREATELY	1	1	Task2	TP	6	6	5	5.75	9	9	9	6	2	6	5	19.75		0.6171875	0.718181818	0.663866	0.526666667	0.473333
					FP	0	0	1	0.25	6	8	6.5	0	4	0	1	7.75						
					FN	1	1	1	1	7	7	7	2	2	2	2	10						
			2		TP	6	6	5	5.75	13	8	11.75	6	2	3	4.25	21.75		0.6796875	0.6796875	0.813084	0.587838	0.412162
					FP	2	2	3	2.25	3	8	4.25	2	6	5	3.75	10.25						
					FN	0	0	0	0	3	3	3	2	2	2	2	5						
			3		TP	6	6	5	5.75	13	8	11.75	6	2	3	4.25	21.75		0.6796875	0.625899281	0.790909	0.537037	0.462963
					FP	3	3	4	3.25	5	9	6	2	6	5	3.75	13						
					FN	0	0	0	0	3	6	3.75	2	2	2	2	5.75						
			4		TP	5	5	4	4.75	11	9	10.5	3	2	3	2.75	18		0.5625	0.774193548	0.6	0.510638	0.489362
					FP	3	3	5	3.5	1	3	1.5	0	1	0	0.25	5.25						
					FN	2	2	2	2	5	5	5	5	5	5	5	12						
			7		TP	7	7	6	6.75	13	12	12.75	7	3	4	5.25	24.75		0.7734375	0.692307692	0.798387	0.589286	0.410714
					FP	0	0	1	0.25	9	9	9	0	4	3	1.75	11						
					FN	1	1	1	1	4	5	4.25	1	1	1	1	6.25						
			8		TP	7	7	6	6.75	14	11	13.25	5	0	2	3	23		0.71875	0.836363636	0.821429	0.707692	0.292308
					FP	0	0	1	0.25	1	4	1.75	1	5	3	2.5	4.5						
					FN	0	0	0	0	2	2	2	3	3	3	3	5						
			9																				

Table D.1: Quality Metrics (Cont)

Quality Metrics																			
Tool	Group	Team	Task	Value	Number of the clase				Number of the attribute				Number of the relations				Element	Succes	Precision	Recall	Accuracy	Error
					Existence	Name	Type	Value	Existence	Type	Value		Existence	Type	cardinalid	Value	Value					
CREATELY	2	14	Task1	TP	6	6	4	5.5	15	9	13.5		5	3	5	4.5	23.5	0.758064516	0.494736842	0.77686	0.433179724	0.56682
				FP	3	3	5	3.5	17	17	17		3	5	3	3.5	24					
				FN	1	1	1	1	3	6	3.75		2	2	2	2	6.75					
		11		TP	5	5	4	4.75	12	10	11.5		4	2	4	3.5	19.75	0.637096774	0.918604651	0.642276	0.607692308	0.392308
				FP	0	0	1	0.25	1	3	1.5		0	0	0	0	1.75					
				FN	2	2	2	2	6	6	6		3	3	3	3	11					
		12		TP	7	5	6	6.55	13	0	9.75		5	3	4	4.25	20.55	0.662903226	0.708620690	0.646226	0.510559006	0.489441
				FP	2	4	3	2.45	4	0	3		3	3	3	3	8.45					
				FN	0	0	0	0	5	22	9.25		2	2	2	2	11.25					
		13		TP	6	6	5	5.75	15	13	14.5		5	0	5	3.75	24					
				FP	0	0	1	0.25	3	5	3.5		0	0	0	0	3.75	0.774193548	0.864864865	0.8	0.711111111	0.288889
				FN	1	1	1	1	3	3	3		2	2	2	2	6					
		14		TP	6	6	5	5.75	17	13	16		5	3	5	4.5	26.25	0.846774194	0.875	0.867769	0.772058824	0.227941
				FP	0	0	1	0.25	2	6	3		0	2	0	0.5	3.75					
				FN	1	1	1	1	1	1	1		2	2	2	2	4					
		15		TP	6	5	4	5.4	8	1	6.25		5	3	3	4	15.65	0.504838871	0.652083333	0.546248	0.422972973	0.577027
				FP	0	1	2	0.6	5	12	6.75		0	2	2	1	8.35					
				FN	1	1	1	1	10	10	10		2	2	2	2	13					
		16		TP	5	5	4	4.75	16	7	13.75		4	2	4	3.5	22	0.709677419	0.8	0.758621	0.637681159	0.362319
				FP	0	0	1	0.25	2	13	4.75		0	2	0	0.5	5.5					
				FN	2	2	2	2	2	2	2		3	3	3	3	7					
		17		TP	6	5	5	5.65	12	8	11		1	0	1	0.75	17.4	0.561290323	0.386666667	0.635036	0.316363636	0.683636
				FP	3	4	4	3.35	14	18	15		9	10	9	9.25	27.6					
				FN	1	1	1	1	4	4	4		5	5	5	5	10					
		18		TP	6	6	5	5.75	10	13	10.75		4	3	4	3.75	20.25	0.653225806	0.5625	0.627907	0.421875	0.578125
				FP	2	2	3	2.25	12	9	11.25		2	3	2	2.25	15.75					
				FN	1	1	1	1	8	8	8		3	3	3	3	12					
SOCIO	2	10	Task2	TP	6	6	5	5.75	15	9	13.5		4	4	3	3.75	23	0.71875	0.691729323	0.754098	0.564417178	0.435583
				FP	1	1	2	1.25	8	11	8.75		0	0	1	0.25	10.25					
				FN	1	1	1	1	1	7	2.5		4	4	4	4	7.5					
		11		TP	7	6	6	6.75	13	9	12		5	5	5	4.5	23.25	0.726562500	0.801724138	0.794872	0.664285714	0.335714
				FP	1	1	2	1.25	0	4	1		3	3	3	3.5	5.75					
				FN	0	0	0	0	3	3	3		3	3	3	3	6					
		12		TP	6	6	5	5.75	12	12	12		5	3	5	4.5	22.25	0.695312500	0.89	0.735537	0.674242424	0.325758
				FP	1	1	2	1.25	0	0	0		1	1	1	1.5	2.75					
				FN	1	1	1	1	4	4	4		3	3	3	3	8					
		13		TP	6	6	5	5.75	12	13	12.25		5	6	6	5.75	23.75	0.742187500	0.848214286	0.772358	0.678571429	0.321429
				FP	0	0	1	0.25	4	3	3.75		0	1	0	0.25	4.25					
				FN	1	1	1	1	4	3	3.75		2	3	2	2.25	7					
		14		TP	6	6	5	5.75	14	11	13.25		4	3	4	3.75	22.75	0.710937500	0.875	0.764706	0.689393939	0.310606
				FP	1	1	2	1.25	1	4	1.75		0	1	0	0.25	3.25					
				FN	1	1	1	1	2	2	2		4	4	4	4	7					
		15		TP	6	6	5	5.75	14	12	13.5		6	4	5	5.25	24.5	0.765625	0.859649123	0.830508	0.731343284	0.268657
				FP	1	1	2	1.25	1	3	1.5		1	1	2	1.25	4					
				FN	1	1	1	1	2	2	2		2	2	2	2	5					
		16		TP	7	7	6	6.75	12	8	11		7	6	6	6.5	24.25	0.75781250	0.734848485	0.776	0.60625	0.39375
				FP	2	2	2	2	5	2	4.25		2	3	3	2.5	8.75					
				FN	0	0	0	0	6	6	6		1	1	1	1	7					
		17		TP	6	6	6	6	13	12	12.75		5	5	5	5	23.75	0.742187500	0.818965517	0.772358	0.659722222	0.340278
				FP	1	1	1	1	1	2	1.25		3	3	3	3	5.25					
				FN	1	1	1	1	3	3	3		3	3	3	3	7					
		18		TP	6	6	6	6	13	13	13		6	3	5	5	24	0.75	0.780487805	0.8	0.653061224	0.346939
				FP	2	2	2	2	2	2	2		2	5	2	2.75	6.75					
				FN	1	1	1	1	3	3	3		2	2	2	2	6					

APPENDIX E
PARTICIPANTS' PREFERENCE

Table E.1: Preference

Group	Team	Preference		Group Preference
		CREATELY	SOCIO	
1	1	1	2	SOCIO
1	2	0	3	SOCIO
1	3	2	1	CREATELY
1	4	2	1	CREATELY
1	5	1	2	SOCIO
1	6	1	2	SOCIO
1	7	1	2	SOCIO
1	8	2	1	CREATELY
1	9	2	1	CREATELY
2	10	1	2	SOCIO
2	11	1	2	SOCIO
2	12	1	2	SOCIO
2	13	2	1	CREATELY
2	14	1	2	SOCIO
2	15	0	3	SOCIO
2	16	0	3	SOCIO
2	17	0	3	SOCIO
2	18	2	1	CREATELY
Total		20	34	

APPENDIX F
TASK DESCRIPTIONS

Appendix F details the task descriptions.

EQUIPO ___

HORA DE INICIO DE LA LECTURA _____
HORA DE FINALIZACIÓN DE LA LECTURA _____

HORA DE INICIO DE LA TAREA _____
HORA DE FINALIZACIÓN DE LA TAREA _____

TAREA 1

Una tienda solicita una aplicación para gestionar sus productos y sus clientes. Disponen de tres tipos de productos: ropa, zapatos y bolsos. Todos los productos tienen identificador, nombre, color, descripción, precio y categoría. En algunas temporadas, los productos pueden tener descuento. La ropa y los zapatos tienen talla, y los zapatos pueden tener diferentes alturas. La tienda desea visualizar toda esta información sobre sus productos, y también, una foto y la cantidad de unidades de los mismos.
La tienda posee los siguientes datos sobre sus clientes: nombre, dirección y número de teléfono. Cada cliente tiene asignado un identificador. Los clientes pueden hacer pedidos. La tienda desea poder registrar los pedidos de cada cliente en la aplicación, para así poder ver la fecha en la que se efectuó el pedido, su identificador y los productos que contiene.

EQUIPO ___

HORA DE INICIO DE LA LECTURA _____
HORA DE FINALIZACIÓN DE LA LECTURA _____

HORA DE INICIO DE LA TAREA _____
HORA DE FINALIZACIÓN DE LA TAREA _____

TAREA 2

Un colegio, cuyo nombre y dirección son conocidos, solicita una aplicación para organizar a sus profesores, alumnos y asignaturas. El colegio imparte asignaturas diferentes en función del curso académico. Cada asignatura consta de varios temas que se podrán gestionar desde la aplicación. Para evaluar cada asignatura se realizan exámenes. El colegio desea poder especificar las preguntas, la fecha y el peso de un examen en la asignatura mediante la app. Se imparten varias clases por asignatura, en un aula, día y hora concretos. Cada clase tiene varios estudiantes y la imparte un solo profesor. El colegio dispone del nombre completo, la dirección, el número de teléfono y la fecha de nacimiento tanto de profesores como de alumnos. Además, toda persona perteneciente al colegio tiene asignado un identificador.

APPENDIX G
QUESTIONNAIRES

Appendix G shows questionnaire SUS for Task 1 and 2 and familiarity questionnaire.

Questionnaire SUS for Task 1/2

GRUPO

HERRAMIENTA

Instrucciones: Para las siguientes afirmaciones, marca la casilla que mejor describa tus reacciones a la herramienta.

	← Strongly disagree				Strongly agree →
Creo que me gustaría usar esta herramienta con frecuencia.....	☐	☐	☐	☐	☐
Encontré esta herramienta innecesariamente compleja.....	☐	☐	☐	☐	☐
Creo que la herramienta es fácil de usar.....	☐	☐	☐	☐	☐
Creo que necesitaría ayuda para poder usar esta herramienta.....	☐	☐	☐	☐	☐
He encontrado que las diversas funciones de esta.....	☐	☐	☐	☐	☐
herramienta estaban bien integradas.....	☐	☐	☐	☐	☐
Creo que hay demasiadas funciones inconsistentes en esta herramienta....	☐	☐	☐	☐	☐
Creo que la mayoría de las personas pueden aprender a usar esta herramienta muy rápidamente.....	☐	☐	☐	☐	☐
He encontrado esta herramienta muy incómoda de usar.....	☐	☐	☐	☐	☐
Me sentí muy seguro de lo que estaba haciendo al usar esta herramienta..	☐	☐	☐	☐	☐
Tengo que aprender muchas cosas antes de poder utilizar esta herramienta.....	☐	☐	☐	☐	☐

Por favor, indica tres aspectos positivos que quieras resaltar sobre la herramienta:

_____

_____

Por favor, indica tres aspectos negativos de la herramienta:

_____

_____

¿Tienes alguna sugerencia de mejora?:

_____

¿Qué herramienta prefieres? (only for Task 2)

SOCIO ☐ CREATELY ☐

Familiarity Questionnaire

GRUPO ___

Cuestiones generales: Para cada una de las siguientes cuestiones rellena o marca la casilla correspondiente.

Edad. . . . . ☐

Sexo. . . . . Hombre ☐ Mujer ☐

¿Eres estudiante o graduado en informática? . . . . . Sí ☐ No ☐

¿Has utilizado alguna vez Telegram?. . . . . Sí ☐ No ☐

¿Has utilizado alguna vez un chatbot?. . . . . Sí ☐ No ☐

¿Qué redes sociales sueles utilizar?. . . WhatsApp ☐ Telegram ☐ Twitter ☐ Facebook ☐ Instagram ☐

Puntúa tu grado de uso de las redes sociales (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Puntúa tu grado de uso de Telegram (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Puntúa tu nivel de inglés (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu grado de conocimiento sobre diagramas de clases (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu nivel de conocimiento sobre chatbots (1-novato, 5-experto). 1 2 3 4 5

Puntúa tu grado de uso de chatbots (1-poco/ninguno, 5-intensivo). 1 2 3 4 5

Versión de Telegram empleada durante la sesión: . . . . . SmartPhone o Tablet ☐ Web ☐ Escritorio ☐

APPENDIX H
IDEAL CLASS DIAGRAMS

Figure H.1: Ideal Class Diagram of Task 1

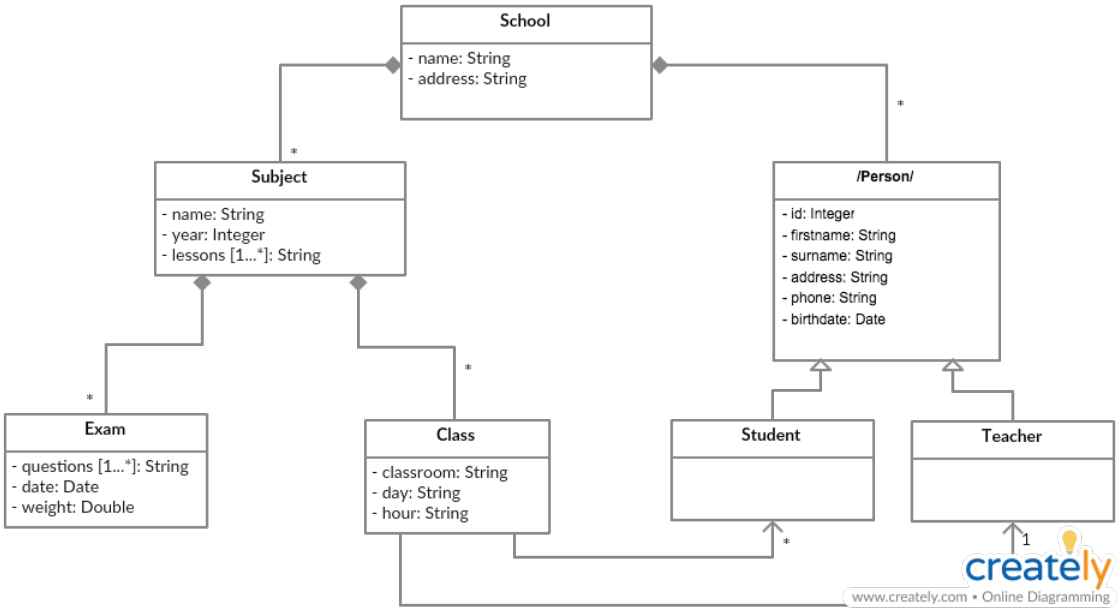

Figure H.2: Ideal Class Diagram of Task 2